Humboldt-Universität zu Berlin
Institut für Bibliothekswissenschaft

Berliner Handreichungen
zur Bibliothekswissenschaft

Heft 126

Entwurf und prototypische Implementierung eines Metadaten-Rahmenwerks für die Digitalisierung an der Universitätsbibliothek Regensburg

von
Helge Knüttel

Entwurf und prototypische Implementierung eines Metadaten-Rahmenwerks für die Digitalisierung an der Universitätsbibliothek Regensburg

von
Helge Knüttel

Berliner Handreichungen
zur Bibliothekswissenschaft

Begründet von Peter Zahn
Herausgegeben von
Konrad Umlauf
Humboldt-Universität zu Berlin

Heft 126

Knüttel, Helge

Entwurf und prototypische Implementierung eines Metadaten-Rahmenwerks für die Digitalisierung an der Universitätsbibliothek Regensburg / von Helge Knüttel. – Berlin : Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin, 2005. – viii, 47 S. – (Berliner Handreichungen zur Bibliothekswissenschaft;126)

ISSN 1438-7662

Abstract:

An der Universitätsbibliothek Regensburg werden in mittlerem Umfang Digitalisate erstellt. Bislang fehlte für die Digitalisate eine Erschließung mit deskriptiven, administrativen und strukturellen Metadaten, die in geeigneter Weise die Benutzbarkeit und die Langzeitarchivierung der digitalen Dokumente unterstützt hätte. Eine vergleichende Untersuchung bibliothekarischer Metadatenmodelle und -formate für die Digitalisierung ergab die besondere Eignung des Metadata Encoding and Transmission Standard (METS) als Metadatenformat für die Digitalisierung an der Universitätsbibliothek Regensburg. Der METS unterstützt ausdrücklich die Funktionalitäten des Referenzmodells eines Open Archival Information System (OAIS) und läßt sich durch standardisierte Erweiterungen an zukünftige Bedürfnisse anpassen. Die neuartige Erschließung ließ sich problemlos in den Geschäftsgang integrieren.

Ausgehend vom METS implementierte ich prototypisch ein Java-basiertes Erschließungswerkzeug, das komplett modular gestaltet ist, um größtmögliche Flexibilität für die Verarbeitung unterschiedlicher Dokumente und Dateiformate sowie für zukünftige Erweiterungen zu erreichen. Das Programm ist plattformunabhängig und als Client-Applikation konzipiert. Es soll die Erschließung der Digitalisate soweit wie möglich automatisieren.

Bei der Planung des Metadaten-Rahmenwerks erwies sich ein System zur Vergabe von Persistent Identifiern (PIDs) als notwendig. PIDs werden für CD-ROMs, METS-Dokumente und Dateien benötigt. Beim Übergang von der Speicherung der digitalen Dokumente auf CD-ROMs zu einem Online-System können die CD-ROM-PIDs entfallen. Das implementierte System zur Vergabe von PIDs kann jederzeit leicht zu einem URN-System mit Resolver weiterentwickelt werden. Neben der PID-Vergabe dient es auch der Authentizitätsprüfung für die digitalen Dokumente.

Ich demonstrierte die mögliche Erzeugung von Präsentationsformen der digitalen Dokumente, mittels XSLT aus den in METS-Dokumenten kodierten Metadaten. Dadurch ließ sich die Benutzbarkeit der Digitalisate deutlich verbessern.

Diese Veröffentlichung geht zurück auf eine Master-Arbeit im postgradualen Fernstudiengang Master of Arts (Library and Information Science) an der Humboldt-Universität zu Berlin.

Diese Arbeit ist im WWW verfügbar unter der URL:
`http://www.ib.hu-berlin.de/~kumlau/handreichungen/h126/`

Inhaltsverzeichnis

Inhaltsverzeichnis **v**

Abkürzungen **vii**

1 Einleitung **1**
1.1 Digitalisierung an der Universitätsbibliothek Regensburg 1
1.2 Problemstellung 3
1.3 Anforderungen an ein Metadaten-Rahmenwerk 4

2 Konzeptioneller Entwurf **7**
2.1 Vergleich bestehender Metadatenstandards 8
2.1.1 Cedars-Projekt 8
2.1.2 NEDLIB-Projekt 8
2.1.3 National Library of Australia 9
2.1.4 OCLC/RLG Preservation Metadata 9
2.1.5 DLmeta 9
2.1.6 Electronic Binding DTD (Ebind) 10
2.1.7 Making of America II (MOA2) 10
2.1.8 Metadata Encoding and Transmission Standard (METS) 11
2.2 Auswahl eines Metadatenstandards 13
2.3 Sicherung der Authentizität von Dokumenten 14
2.3.1 Abstammung vom analogen Original 14
2.3.2 Digitale Derivate 16
2.3.3 Authentizität 17
2.4 Persistente Identifier 18

3 Prototypische Implementierung **21**
3.1 Programmierung eines Erschließungswerkzeugs 21
3.1.1 Ablauf der Erschließung 22
3.1.2 Dateien, technische und strukturelle Metadaten 23
3.1.3 Bibliographische Metadaten 24
3.1.4 Graphische Benutzeroberfläche 24
3.1.5 Erzeugung von Präsentationsformen 25
3.2 Programmierung eines Systems für Persistente Identifier 25
3.2.1 Persistent Identifier für CD-ROMs 26
3.2.2 Persistent Identifier für METS-Dokumente 26
3.2.3 Persistent Identifier für Dateien 27
3.3 Organisatorische Veränderungen bei der Digitalisierung 27

3.3.1 Bisheriger Geschäftsgang 27
3.3.2 Neuerungen 28
3.4 Hardwareanforderungen 29

4 Exemplarische Erschließung von Digitalisaten 31
4.1 Komplette Erschließung eines Einzelbildes 31
4.2 Digitalisat eines ganzen Bandes mit Strukturinformationen 33
4.3 Erzeugung eines Inhaltverzeichnisses 35

5 Diskussion und Bewertung 39

Literaturverzeichnis 41

Danksagung 47

Abkürzungen

AIP Archival Information Package; siehe OAIS (2002).

BVB BibliotheksVerbund Bayern

CD-ROM Compact Disc - Read Only Memory

CGI Common Gateway Interface

DDB Die Deutsche Bibliothek

DIP Dissemination Information Package; siehe OAIS (2002).

DTD Document Type Definition

DVD Digital Versatile Disc

HTML Hypertext Markup Language[1]

JPEG Joint Photographic Experts Group[2]

MAB Maschinelles Austauschformat für Bibliotheken

METS Metadata Encoding and Transmission Standard; siehe METS (2003).

MIME Multipurpose Internet Mail Extensions; siehe Freed und Borenstein (1996).

MIX / NISO MIX NISO Metadata for Images in XML; siehe MIX-Website (2003).

NISO National Information Standards Organization[3]

OAIS Open Archival Information System; siehe OAIS (2002).

OCR Optical Character Recognition

OPAC Online Public Access Catalog

pdf Portable Document Format[4]

PID Persistent Identifier

SGML Standard Generalized Markup Language

[1] Siehe `http://www.w3.org/MarkUp/`
[2] Siehe `http://www.jpeg.org/`
[3] Siehe `http://www.niso.org/index.html`
[4] Siehe `http://partners.adobe.com/asn/tech/pdf/specifications.jsp`

SIP Submission Information Package; siehe OAIS (2002).

TIFF Tagged Image File Format; auch mit Versionsangabe: TIFF4, TIFF6 (TIFF 6.0 1992), TIFF/EP (NORM ISO 12234-2:2001 2001).

URI Uniform Resource Identifier; siehe Berners-Lee u. a. (1998).

URL Uniform Resource Locator; siehe Berners-Lee u. a. (1994).

URN Uniform Resource Name; siehe Moats (1997).

XML Extensible Markup Language; siehe XML (2004).

XSL Extensible Stylesheet Language

XSLT XSL Transformations; siehe XSLT (1999).

1

Einleitung

Die Metadaten, die für ein erfolgreiches Management und eine gute Benutzbarkeit digitaler Objekte notwendig sind, sind sowohl umfangreicher als auch verschieden von den Metadaten, die traditionell für die Verwaltung gedruckter Werke verwendet werden (METS 2003). Wenn eine Bibliothek deskriptive Metadaten über ein Buch in ihrem Bestand vorhält, zerfällt das Buch nicht in eine Reihe einzelner, unverbundener Seiten, wenn die Bibliothek es unterläßt, Metadaten zu seiner Struktur zu erfassen. Genausowenig wird es für einen Wissenschaftler unmöglich, den Wert eines Buches einzuschätzen, wenn die Bibliothek nicht vermerkt hat, auf welcher Druckerpresse das Buch hergestellt worden ist. Dasselbe gilt aber nicht für die digitalisierte Version des Buches. Ohne Strukturmetadaten sind die Einzelbilder der Seiten oder die Textdateien, die das Werk ausmachen, von geringem Wert und ohne technische Metadaten zum Digitalisierungsprozeß sind Wissenschaftler unter Umständen unsicher, wie exakt das Original im digitalen Dokument wiedergegeben wird. Für Zwecke der internen Verwaltung muß eine Bibliothek zudem Zugriff auf angemessene technische Metadten haben, um die Daten periodisch erneuern oder migrieren zu können und damit die langfristige Verfügbarkeit der wertvollen Ressourcen sicherstellen zu können (METS 2003).

1.1 Digitalisierung an der Universitätsbibliothek Regensburg

Die Universitätsbibliothek Regensburg besitzt seit Beginn des Jahres 2002 im Rahmen des Multimediazentrums[1] Digitalisierungsmöglichkeiten. Für Benutzer frei zugänglich sind acht DIN-A4-Flachbettscanner und zwei Dia-/Filmscanner an PCs mit CD-Brennern im Bibliographiensaal der Zentralbibliothek. Hier können Benutzer kostenlos, aber ohne Beratung für eigene Bedürfnisse scannen.

Daneben gibt es auch einige hochwertigere Geräte in den Räumen des Multimediazentrums, die von Benutzern nur unter Aufsicht und nach einer Einweisung benutzt werden können. Hier werden auch gegen geringe Gebühr Digitalisierungsaufträge für Benutzer erledigt und, teilweise im Rahmen drittmittelfinanzierter Projekte, Digitalisate für den Bestand der Universitätsbibliothek Regensburg erstellt.

Derzeit stehen dort folgende Geräte für die Digitalisierung zur Verfügung:

[1] `http://www.bibliothek.uni-regensburg.de/mmz/indexmmz.html`

- Ein A0-Großformat-Scanner ProServ TriAS. Er eignet sich besonders für verzerrungsfreie Farbscans großer Vorlagen. Im Projekt Bayerische Landesbibliothek Online[2] werden hier ca. 2000 historische Landkarten Bayerns aus dem Bestand der Bayerischen Staatsbibliothek digitalisiert und nach einer Georeferenzierung auf dem Zoomserver der Bayerischen Staatsbibliothek im WWW bereitgestellt. Ausgehend von einer Suche in der Ortsdatenbank können über die Gauß-Krüger-Koordinaten des Ortes in einem Geographischen Informationssystem Karten ermittelt werden, die diesen Punkt enthalten.

- Ein Graustufen-Buchscanner Zeutschel OS 5000. Neben dem Einsatz für die Fernleihe und für den Dokumentenlieferdienst subito[3] werden hier vor allem auf Benutzerwunsch schonend ältere Monographien digitalisiert, die dann teilweise in digitaler Form in den Bestand aufgenommen werden.

- Ein Scanner Canon MicrofilmScanner 800 für Mikrofilme und Mikrofiches. Von Benutzern wird die Digitalisierungsmöglichkeit von Mikrofilmen und Mikrofiche (v. a. Dissertationen und Musikalia) sehr gerne angenommen, da sie viel schnellere und preiswertere Kopien erlaubt, als der früher verwendete Readerprinter.

- Ein hochwertiger DIN-A3-Flachbettscanner Epson Expression 1640 XL. Hier werden v. a. farbig illustrierte, ältere Monographien, kleinere Karten und Einblattdrucke digitalisiert, um sie in digitaler Form für die Benutzung verfügbar zu machen. Aus dem Bestand der Bibliothek der Regensburgischen Botanischen Gesellschaft[4], der als Dauerleihgabe an der Universitätsbibliothek Regensburg ist, wird zur Zeit das mit 2472 kolorierten Kupfertafeln illustrierte Werk von Sturm (1797–1862) digitalisiert, das in der vorliegenden ersten Auflage zu den gesuchtesten illustrierten botanischen Werken gehört. Es soll nach der Digitalisierung im WWW zur Nutzung angeboten werden.

- Ein Audio-Arbeitsplatz mit Verstärker, Plattenspieler, Tonbandgerät, Cassetten-Deck und der Digitalisierungsmöglichkeit durch eine hochwertige PC-Soundkarte Terratec DMX 6fire 24/96. An diesem Arbeitsplatz werden derzeit vor allem die ca. 7900 analogen Magnettonbänder des Historischen Werbefunk-Archivs von Prof. E. H. Geldmacher, Honorarprofessor an der Universität der Künste Berlin, digitalisiert, die der Universitätsbibliothek Regensburg überlassen wurden. Die Dauer dieses Projekts wird ca. zwei Jahre betragen. Dabei soll der Bestand der teilweise 50 Jahre alten Tonbänder gesichert und die Nutzung für die Forschung wesentlich verbessert werden.

Durch das Vorhandensein dieser Geräte kann Benutzern teilweise die schonende Digitalisierung benötigter Literatur angeboten werden, die sonst nur im Handschriftenlesesaal genutzt werden könnte. Das Digitalisat kann der Benutzer hingegen auf CD-ROM mitnehmen. Die digitalisierten, urheberrechtsfreien Werke werden in den Bestand der Universitätsbibliothek Regensburg übernommen und stehen Benutzern zukünftig direkt zur Verfügung. Das Original muß seltener eingesehen werden, wird also, ganz im Sinne der Bestandserhaltung, geschont.

Daneben werden urheberrechtsfreie Medien auch unabhängig von aktuellen Benutzerwünschen für den Bestand der Bibliothek digitalisiert. Dies geschieht teilweise für den

[2] http://mdz2.bib-bvb.de/~blo/

[3] http://www.subito-doc.de/

[4] http://www.regensburgische-botanische-gesellschaft.de/

Bestandserhalt gefährdeter Werke, vor allem aber, um seltene und alte Werke der Benutzung zugänglich machen zu können. Ende des Jahres 2003 standen den Benutzern so 113 urheberrechtsfreie, digitalisierte Werke zur Ausleihe auf CD-ROM zur Verfügung.

Die Digitalisierung und ihre Möglichkeiten sind an der Universitätsbibliothek Regensburg historisch gewachsen. Während die Digitalisierng anfangs vor allem für die direkte Benutzung angeboten wurde, wurden dann mit der besseren Infrastruktur (mehr und bessere Scanner, Audioarbeitsplatz) immer mehr Digitalisate erstellt und schließlich kamen auch ganze Digitalisierungsprojekte dazu. Für einzelne Projekte wurden dabei jeweils bedarfsgerechte deskriptive Metadaten insbesondere bibliographischer Natur erhoben (z. B. Sammlung von Porträts des Hauses Thurn und Taxis[5]). Bisher standen nie, auch nicht aus Projektmitteln, die Ressourcen zur Verfügung, um ein allgemeines Konzept für die organisierte Metadatenerfassung bei der Digitalisierung zu erarbeiten und dann auch umzusetzen. Dies gilt besonders für technische Metadaten als Basis der Langzeitarchivierung und für Strukturmetadaten.

Metadaten wurden bei der Digitalisierung bisher nur minimal erstellt. Die Struktur digitaler Dokumente spiegelt sich im allgemeinen in den Namen der erstellten Dateien wieder. Die Benennung der Dateien wurde formal nie festgelegt, sondern war einer gewissen Evolution mit zunehmender Erfahrung des Personals unterworfen, erfolgte aber immer nach praktischen Gesichtspunkten am jeweiligen Werk. Die Dateien wurden meist durchnumeriert, um die Reihenfolge, z. B. die Seitenfolge, anzugeben. Teilweise wurden auch zusätzliche Informationen in den Dateinamen kodiert, beispielsweise unterschiedlichen Paginierungen, Tafelseiten oder bei manchen Werken auch weitergehende inhaltliche Informationen. Während sich dieses Prinzip bei der täglichen Arbeit durchaus bewährt hat, ist es dennoch nicht standardisiert und teilweise ohne Erläuterung nicht unmittelbar zu verstehen.

Die Archivierungsform der Digitalisate diente bisher auch als Präsentationsform. Dies sind CD-ROMs mit den Rohdaten (z. B. TIFF-Dateien) aus der Digitalisierung ohne weiteres verbindendes Element als der Lage in einem gemeinsamen Verzeichnis im Dateisystem. Erst in jüngerer Zeit wurde bei geeigneten Vorlagen zusätzlich eine pdf-Datei erstellt, die die Einzelbilder zusammenfaßt und damit das Blättern und Ausdrucken erleichtert.

Den Digitalisaten wurden teilweise einige Metadaten in einer Textdatei auf der CD-ROM beigegeben. Eine Kopie der Katalogaufnahme der analogen Quelle aus dem Verbundkatalog des BVB stellte die bibliographischen Metadaten dar. Dazu kamen minimale Angaben zum Dateiformat und zum verwendeten Scanner. Bei Werken, die wegen ihrer Größe in digitalisierter Form auf mehrere CD-ROMs gebrannt werden mußten, wurde so eine Textdatei aber nur auf der ersten CD-ROM gespeichert. Abgesehen von der äußeren Beschriftung gab es keine Angaben auf den CD-ROMs, die ihre Zusammengehörigkeit erkennen liessen oder, abgesehen von der ersten CD-ROM, die Herkunft vom analogen Original anzeigten. Der bisherige Zustand wird allgemein als unbefriedigendes Übergangsstadium angesehen.

1.2 Problemstellung

In dieser Arbeit soll ein Metadaten-Rahmenwerk für die Digitalisierung an der Universitätsbibliothek Regensburg konzipiert und prototypisch implementiert werden. Gegenstand der Arbeit sind dabei nur die an der Universitätsbibliothek Regensburg erzeugten Digitalisate, die auch in den Bestand aufgenommen werden. Diese haben bisher schon, entweder als

[5] http://www.bibliothek.uni-regensburg.de/projekte/portraets/portraets.htm

selbständige oder unselbständige Werke, einen Eintrag im Bibliothekskatalog. Sie werden also bibliothekarisch voll erschlossen oder sie werden als Teil eines Sammelwerks objektspezifisch mit deskriptiven Metadaten erschlossen. Der zweite Fall trifft oft auf Medien aus Sammlungen zu, die im Rahmen eines drittmittelfinanzierten Projekts digitalisiert werden. Für diese Metadaten wurden meist eigene Datenbanken geschaffen, die den Besonderheiten der Objekte Rechnung tragen (z. B. Regensburger Porträtgalerie mit digitalisierten Graphiken des Hauses Thurn und Taxis[6], historische Karten von Bayern im Rahmen der Bayerischen Landesbibliothek Online[7]).

Was bisher vor allem fehlt, sind Strukturmetadaten, technische und administrative Metadaten für die Digitalisate, sowie ein Rahmen, der diese zusammen mit den schon vorhandenen bibliographischen Metadaten zu einem benutzbaren Ganzen integriert.

Diese zusätzlichen Metadaten werden zum einen benötigt, um die bei der Digitalisierung entstandenen Dateien überhaupt zu Einheiten zusammenzufassen, die ihre Herkunft vom Original kennzeichnen. Besonders notwendig ist dies für die Schaffung von Präsentationsformen für die Benutzung, die bisher noch nicht vorhanden sind. Desweiteren ist auch die Möglichkeit für eine tiefere Erschließung der digitalen Dokumente wünschenswert, um deren Benutzung zu erleichtern. Eine ganz wesentliche Aufgabe der Metadaten ist es außerdem, den Zugang zu den Inhalten auf längere (bestenfalls unbestimmte) Zeit zu gewährleisten, in dem sie die digitale Langzeitarchivierung in geeigneter Weise unterstützen.

Neben der Auswahl entsprechender Metadatenformate und der Schaffung von Softwarewerkzeugen für die Erschließung müssen auch die Einbindung der neuartigen Erschließung in den Geschäftsgang der Digitalisierung und Hardwareanforderungen konzipiert werden. Diese notwendige Kombination aus technischen und organisatorischen Aspekten bezeichne ich als Metadaten-Rahmenwerk. Die Bezeichnung Rahmenwerk hat aber noch eine weitere Begründung. Im Feld der digitalen Langzeitarchivierung sind noch sehr viele Probleme ungelöst; Best-Practice-Empfehlungen und Standards befinden sich in der aktiven Entwicklung. Zum Zeitpunkt dieser Arbeit kann deshalb kein abschließendes und dauerhaft gültiges Metadatenmodell mit einer fixen Implementierung abgeliefert werden. Vielmehr will ich einen Rahmen für die Metadatenerhebung schaffen, der bezüglich der Metadatenelemente, ihrer semantischen und syntaktischen Struktur, des Geschäftsgangs und der Softwarewerkzeuge flexible Anpassungen an sich neu entwickelnde Anforderungen, Dateiformate und Standards erlaubt. Dieser Rahmen soll sofort benutzbare Komponenten enthalten, aber zugleich ausbaufähig für die Zukunft sein. Zentrale Voraussetzung dafür ist ein geeignetes Metadatenformat.

Um die Praxistauglichkeit der hier vorgestellten Konzepte zu prüfen, werden Softwarewerkzeuge im Sinne eines *proof of concept* prototypisch entwickelt, die als Bausteine für eine spätere, vollständige Implementierung dienen können.

1.3 Anforderungen an ein Metadaten-Rahmenwerk für die Digitalisierung

Die Anforderungen an ein Metadaten-Rahmenwerk für die Digitalisierung an der UBR ließen sich sehr knapp zusammenfassen: Die Metadaten müssen die Erstellung von Präsentationsformen und vor allem die digitale Langzeitarchivierung in geeigneter Weise unterstüt-

[6] http://www.bibliothek.uni-regensburg.de/projekte/portraets/portraets.htm
[7] http://mdz2.bib-bvb.de/~blo/

zen. Unter digitaler Langzeitarchivierung verstehe ich dabei die Gewährleistung der dauerhaften Zugänglichkeit eines digitalen Dokuments. Insofern ist über die Zugänglichkeit auch die Erzeugung von Präsentationsformen für die Benutzung bereits in der Langzeitarchivierung enthalten.

In diesem Zusammenhang ist das „Reference Model for an Open Archival Information System“ (OAIS 2002) von besonderer Bedeutung, das inzwischen zur ISO-Norm erhoben wurde (NORM ISO 14721:2003 2003). Ein *Open Archival Information System* (OAIS) ist definiert als ein „Archiv“, bestehend aus einer Organisation von Menschen und technischen Systemen, das die Aufgabe übernommen hat, Information langfristig zu erhalten und für eine Zielgruppe verfügbar zu machen (“*that has accepted the responsibility to preserve information and make it available to a desginated community*”) (OAIS 2002). Das OAIS-Modell ist ein Modell definierter Verantwortlichkeiten und Funktionen, um diese Aufgabe zu erreichen. Es soll als ein abstraktes Bezugssystem mit klar definierten Begrifflichkeiten dienen, auf das man sich bei der Planung und Implementierung eines Systems mit der genannten Verantwortlichkeit beziehen kann (deshalb Referenzmodell). Ein OAIS unterscheidet sich von anderen „Archiven“ – in OAIS (2002) allgemein verwendet im Sinne von informationsspeichernden und -verwaltenden Institutionen – dadurch, daß es die im OAIS-Modell genannten Verantwortlichkeiten ausdrücklich aufgreift und zu lösen versucht. Wichtig ist mir dabei, die Betonung der geeigneten Aufbereitung der verwalteten Informationen für die Nutzer des Systems hervorzuheben.

Ein OAIS ist ausdrücklich nicht auf den digitalen Bereich beschränkt. Letztlich kann sich jede Bibliothek mit dem Auftrag der Bestandserhaltung (und jedes Museum, Archiv etc.) als ein OAIS verstehen und sollte dies vielleicht auch.

Das Ziel der Erstellung eines Metadaten-Rahmenwerks in dieser Arbeit ist letztlich, mit einer geeigneten Metadatenerhebung die erste Basis für ein OAIS zu gründen. Das Fernziel jenseits dieser Arbeit ist die vollständige Umsetzung dieses Modells, das heißt, ein System zu planen und zu implementieren, das die Verantwortlichkeiten des OAIS-Referenzmodells für die Digitalisate übernimmt und in ein funktionierendes System umsetzt. Für ein OAIS ist vor allem auch die Managementebene wichtig, die regelmäßig und strukturiert Vorgaben macht, wie das Endziel der digitalen Bestandserhaltung zu erreichen ist. Diese organisatorischen Vorgaben können und sollen hier nicht gelöst werden, da die Aufgaben dafür zu umfangreich sind. Die Definition von Strategien und die Schaffung der notwendigen organisatorischen Strukturen für die digitale Bestandserhaltung sind unter Umständen schwieriger als die Lösung technischer Probleme (Day 2001).

Folgende Anforderungen soll das zu schaffende Metadaten-Rahmenwerk im einzelnen erfüllen:

- Das verwendete Rahmenformat für die Metadaten muß sich am OAIS-Referenzmodell orientieren oder ausdrücklich Bezug darauf nehmen.

- Die Metadatenformate müssen unbedingt offen und nicht proprietär sein. Nicht dokumentierte Datenformate, zu denen der Zugang gar noch rechtlich beschränkt ist, lassen nur die größten Probleme bei der Weiterverwendung bis hin zum Datenverlust erwarten. Ich mußte selber bereits entsprechende Erfahrungen mit einem weitverbreiteten kommerziellen Programm machen.

- Das Metadatenformat muß mit bestehenden, internationalen Standards abgestimmt sein bzw. einem solchen Standard entsprechen (siehe auch nächsten Punkt).

- Die Metadatenformate sollten möglichst viele Anwender haben. Eine breite Unterstützung ist sicher mit die beste Gewähr für zukünftige Weiterentwicklungen und eine dauerhafte Zugänglichkeit zu den Inhalten.
- Der Metadatenstandard soll in einem offenen Prozeß entwickelt worden sein. Solche, zumeist durch ausgewiesene Spezialisten aus unterschiedlichen Interessengruppen entstandene Standards haben oft schon eine längere Entwicklungszeit hinter sich. Deshalb können eine bessere Eignung für den Zweck (z. B. technische Beschreibung von Einzelbildern) bei größerer Flexibilität und eine längere Stabilität und Unterstützung eines so entstandenen Standards erwartet werden.
- Im Sinne der Offenheit und Möglichkeit für Weiterentwicklungen sowohl des verwendeten Rahmenstandards, als auch für evtl. darin eingebettete Standards, sollten nach Möglichkeit XML und XML-Schema zum Einsatz kommen. Die Verwendung von XML und verwandten Techniken (z. B. XML-Schema, XSLT) wird als besonders günstig für die langfristige Zugänglichkeit digitaler Dokumente angesehen (Dobratz u. a. 2001).
- Die Metadatenerschließung muß sich ohne allzu große Umstände in die bestehenden Arbeitsabläufe einbetten lassen.
- Der Metadatenstandard muß jetzt in einer für ein Programm verwendbaren Form (z. B. XML-Schema, Softwarewerkzeuge) verfügbar sein.

2

Konzeptioneller Entwurf eines Metadaten-Rahmenwerks

In diesem Abschnitt soll ein Metadaten-Rahmenwerk für die Digitalisierung an der Universitätsbibliothek Regensburg konzipiert werden. Dazu werde ich zuerst verfügbare Metadatenstandards und aktuelle Entwicklungen von Metadaten für Digitalisate vergleichend betrachten (Abschnitt 2.1), um dann aus diesen Alternativen anhand der in Abschnitt 1.3 aufgeführten Anforderungen den am geeignetsten erscheinenden Metadatenstandard auszuwählen (Abschnitt 2.2). Anschließend betrachte ich, wie mit dem gewählten Metadatenstandard die Authentizität (im weiteren Sinne) der digitalen Dokumente gesichert werden kann (Abschnitt 2.3). Persistente Identifier sind ein wesentlicher Bestandteil von Metadaten für digitale Dokumente (Schroeder 2003). Deshalb untersuche ich abschließend, wo und wie diese im hier entworfenen System eingesetzt werden sollten (Abschnitt 2.4). Organisatorische Veränderungen beim Geschäftsgang der Digitalisierung betreffen eher praktische Aspekte und werden deshalb in Abschnitt 3.3 behandelt.

In dieser Arbeit kann ich nur versuchen, konzeptionelle und technische Grundlagen bei den Metadaten der Digitalisate zu schaffen, damit später mit diesen wie beschrieben ein System erstellt werden kann, das OAIS-konform der digitalen Langzeitarchivierung dient. Die Metadaten sollen also zuerst als Grundlage und erster Schritt für spätere Weiterentwicklungen erhoben werden. Die Erhebung standardisierter Metadaten hat darüberhinaus aber auch unmittelbare Vorteile, da sie etwa der Erstellung von Präsentationsformen dienen können und so die Benutzung sehr erleichtern (siehe Abschnitt 4.3, S. 35).

Den weiteren Ausführungen liegt folgendes Konzept zugrunde: Bei der Digitalisierung entsteht aus dem analogen Dokument (z. B. eine Monographie, eine Landkarte oder ein Tonband) ein digitales Dokument. Das entstandene digitale Dokument besteht oft aus mehreren Dateien (z. B. Bilddateien von den einzelnen Seiten einer Monographie), die selber wiederum als digitale Dokumente bezeichnet werden können. Die einzelnen Dateien können ggf. nach der physischen oder logischen Struktur des Werks zu weiteren Einheiten (z. B. Kapiteln einer Monographie) zusammengefaßt werden, die auch als digitale Dokumente aufgefaßt werden können. Eine einem ursprünglichen, als eine Einheit erscheinenden analogen Werk entsprechende digitale Manifestation kann also aus einer ganzen Hierarchie digitaler Dokumente bestehen, deren oberste Stufe in ihrer Granularität dem ursprünglichen Werk entspricht. Aufgabe der Metadaten für Digitalisate ist nun vor allem auch, die einzelnen Dateien gemäß ihrer Zugehörigkeit zum analogen Original zusammenzufassen und dabei, mehr oder weniger tief erschlossen, dessen Struktur wiederzugeben (siehe Abschnitt 2.3.1, S. 14).

Hier wird immer von Dateien als der Basiseinheit eines digitalen Dokuments gesprochen, weil dies der geläufige Begriff für die herkömmliche Verwaltungseinheit digitaler Dokumente ist. Derzeit werden an der Universitätsbibliothek Regensburg die Ergebnisse der Digitalisierung auch tatsächlich als einzelne Dateien in einem Dateisystem (auf Festplatte bzw. CD-ROM) gespeichert. Dies ist konzeptionell aber nicht notwendig und kann sich in Zukunft ändern. Dateien im hier verwendeten Sinne müssen nicht abgeschlossene Einheiten in einem Dateisystem sein, sondern sie können auch als Entitäten eingebunden in eine andere Datei oder ein größeres System vorkommen, etwa in eine XML-Datei oder in einer Multimedia-Datenbank. So können etwa nach dem Metadata Encoding and Transmission Standard (METS; siehe Abschnitt 2.1.8, S. 11) die in den Metadaten beschriebenen digitalen Dokumente (und auch Metadaten, die METS erweitern) entweder direkt in die METS-XML-Datei eingebettet werden oder aber als externe Dateien referenziert werden. Momentan werden aus praktischen Erwägungen die Digitalisate als Dateien referenziert, Metadaten aber eingebunden.

2.1 Vergleich bestehender Metadatenstandards

Im folgenden werde ich die wichtigsten, veröffentlichten Metadatenstandards und -definitionen der letzten Jahre vorstellen, die sich ausdrücklich der digitalen Bestandserhaltung in Bibliotheken widmen. Die Darstellungen sind sehr knapp gehalten, um den Umfang dieser Arbeit nicht zu sprengen. Neben einer kurzen Einordnung in das Umfeld der Standards werden nur die Punkte hervorgehoben, die für meine Zielsetzung bedeutsam sind. Einen Überblick über die Entwicklung von Metadaten für die digitale Bestandserhaltung bis ins Jahr 2001 geben auch Day (2001) und PMFWG (2001).

2.1.1 Cedars-Projekt

Das Cedars-Projekt der Universitäten von Cambridge, Leeds und Oxford (Cedars 2002) lief von 1998 bis 2002. Ziel des Projekts war, strategische, methodologische und praktische Probleme der digitalen Bestandserhaltung zu untersuchen und mögliche Antworten zu finden. Als eines der Ergebnisse des Projekt entstand eine Metadatendefinition, die sich sehr stark an den Entitäten des OAIS-Referenzmodells orientierte (Cedars-Metadaten 2002). Die Elemente waren auf einer recht hohen Ebene definiert, wobei angenommen wurde, daß einige Elemente sinnvollerweise in (nicht weiter definierte) Unterelemente aufgeteilt werden könnten (Day 2001). Neben der Implementierung des Metadatensatzes im Projekt ist mir kein weiterer Einsatz bekannt.

2.1.2 NEDLIB-Projekt

Im Rahmen des NEDLIB-Projekts (Networked European Deposit Library, Laufzeit 1998–2000) (NEDLIB 2000) entstand ebenso wie im Cedars-Projekt ein eng an das OAIS-Modell angelehnter Satz von Metadaten für die digitale Bestandserhaltung (NEDLIB-Metadaten 2000). Dieser wurde als ein Minimalsatz von vorgeschriebenen „Kern-“Elementen technischer Metadaten konzipiert. Administrative, deskriptive und rechtliche Metadaten fehlen. Implementierungen außerhalb des Projekts sind mir nicht bekannt.

2.1.3 National Library of Australia

Die National Library of Australia[1] hat sich seit längerem der digitalen Langzeitarchivierung verpflichtet. Sie publizierte 1999 den Entwurf eines Metadatensatzes für die Bestandserhaltung digitaler Sammlungen (NLA 1999). Die Metadaten waren auf eine dreistufige Hierarchie der beschriebenen Dokumente ausgelegt. Es wurden Sammlungen, Objekte und Dateien unterschieden. Seit 1999 gab es offenbar keine direkte Weiterentwicklung dieses Metadatensatzes.

2.1.4 OCLC/RLG Preservation Metadata

Das Online Computing Library Center OCLC[2] und die Research Libraries Group RLG[3] begannen 2001 mit der gemeinsamen Erarbeitung von Metadaten für die digitale Langzeitarchivierung. Im Jahre 2002 wurde ein Metadatensatz publiziert (PMFWG 2002), der sich am OAIS-Modell orientierte und vor allem auf den Erfahrungen des Cedars-Projekts (siehe Abschnitt 2.1.1), des NEDLIB-Projekts (siehe Abschnitt 2.1.2) und der National Library of Australia (siehe Abschnitt 2.1.3) aufbaute. Der Metadatensatz ist natürlichsprachlich beschrieben. Es liegt bisher noch keine direkt einsetzbare Implementierung, z. B. als XML-Schema, vor. Die Anstrengungen von OCLC/RLG dauern an. So versucht derzeit eine Arbeitsgruppe, einen Satz an „Kernmetadaten“ zu identifizieren, der gemeinsam in den verschiedenen Initiativen zur digitalen Bestandserhaltung vorkommt. Außerdem werden Implementierungsempfehlungen für den Einsatz von Metadaten für die digitale Bestandserhaltung erarbeitet. Ergebnisse dieser Aktivitäten sind noch nicht öffentlich verfügbar.

2.1.5 DLmeta

Das Datenmodell DLmeta (DLmeta 2000) wurde Ende der 1990er-Jahre während mehrerer Projekte in Baden-Württemberg entwickelt und in Form einer XML-DTD (DLmeta-DTD 2000) implementiert (Spanier 2004). Es bietet die Möglichkeit, bibliographische, inhaltliche und strukturelle Metadaten zu (nicht nur digitalen) Objekten zu erfassen. DLmeta baut auf dem Dublin-Core-Metadatensatz (Dublin Core 2003) auf, der um zusätzliche, in den Projekten als notwendig erachtete Elemente erweitert wurde. Im Element `<Local>` können außerdem beliebige Erweiterungen für lokale Bedürfnisse vorgenommen werden (DLmeta-DTD 2000). Die Verarbeitung von DLmeta-Dokumenten in anderen Softwaresystemen wird dadurch aber schwer oder sogar unmöglich gemacht (Spanier 2002). Besser ist der Ansatz von METS (siehe Abschnitt 2.1.8, S. 11) standardisierte Erweiterungen zu verwenden, deren Format ausdrücklich angegeben wird (METS 2003).

In Baden-Württemberg wird DLmeta in mehreren Systemen produktiv eingesetzt (Aschoff u. a. 2004), so im timms-Multimedia-Server[4] (Abele 2004) und der Lokalen Medienbibliothek[5] (Werkmeister und Bognar 2004) der Universität Tübingen. Die Verwendung von DLmeta blieb bisher aber wohl auf dieses Bundesland beschränkt, was auch an der nicht sehr nachdrücklich betriebenen Propagierung des Formats nach außen liegen dürfte. Die

[1] `http://www.nla.gov.au/`
[2] `http://www.oclc.org/`
[3] `http://www.rlg.org/`
[4] `http://www.uni-tuebingen.de/timms`
[5] `http://plautus.ub.uni-tuebingen.de/lmb/opac/`

Website mit der Dokumentation zum DLmeta-Format wurde seit ihrer Gründung im Jahre 2000 nicht fertiggestellt (DLmeta 2000). Das 2002 als Weiterentwicklung entstandene XML-Schema-basierte und deutlich erweiterte Format DLmeta2002 (Spanier 2002, 2004) ist bisher nicht publiziert. Das Format soll auf dem Stand von DLmeta2002 vorerst stabil gehalten werden, wobei ein Einsatz außerhalb von Baden-Württemberg bisher nicht abzusehen ist (Kurt Spanier, persönliche Mitteilung, Jan. 2004).

2.1.6 Electronic Binding DTD (Ebind)

Die Electronic Binding DTD (Ebind) (Ebind-DTD 1996) wurde 1996 an der University of California at Berkeley als eine SGML-DTD entwickelt. Damals begann man, vom Papierzerfall bedrohte Bücher auf einem neuartigen Digitalkopierer zu scannen. Die aus der Massendigitalisierung resultierenden Einzelbilder der Seiten konnten in Ebind-kodierten SGML-Dokumenten, die gemeinsam mit den Bildaten gespeichert wurden, zusammengefaßt („gebunden“) werden. Dabei wurden nur die wichtigsten bibliographischen und Strukturmetadaten erfaßt. Das Metadatenformat war für eine möglichst einfache Erschließung durch Hilfskräfte bei großem Durchsatz konzipiert. Es war nur auf die recht einheitliche Struktur der Dokumente und auf das Format der vom Digitalkopierer erzeugten bitonalen TIFF4-Dateien ausgelegt. Es fehlen technische Metadaten, die für die digitale Langzeitarchivierung unabdingbar sind. Ebind erreichte einige Popularität, vor allem auch wegen der dazu verfügbaren Skripte, mit denen sich sehr einfach HTML-Inhaltsverzeichnisse und HTML-Seiten zum Blättern erzeugen ließen. Ebind wurde in Berkeley schon bald von den Entwicklungen für das MOA2-Format (siehe Abschnitt 2.1.7) abgelöst, die schließlich zum METS (siehe Abschnitt 2.1.8) führten (Ebind 2002).

2.1.7 Making of America II (MOA2)

Im Zentrum der Entwicklung des Projekts Making of America II (MOA2 2001) stand ab 1998 die Entwicklung einer XML-DTD (MOA2-DTD 2000), die als Syntax für den Austausch digitaler Dokumente dienen konnte. Es sollten definierte deskriptive, administrative und Struktur-Metadaten zusammen mit dem primären Inhalt zu einem einheitlichen Objekt kombiniert werden können. Ein solches Objekt mit standardisierter Kodierung sollte dann den Aufbau von Systemarchitekturen ermöglichen, die auf skalierbarer, objektorienterter Softwaretechnologie basieren. Dieser Ansatz wurde im Gegensatz zu der Herangehensweise gesehen, eine „Digitale Bibliothek“ aus CGI-Skripten und HTML-Seiten zusammenzuflicken (MOA2 2001).

Es wurden Softwarewerkzeuge für die Erzeugung, Speicherung und Verwaltung von Metadaten, sowie für die Anzeige der Dokumente entwickelt. Ziel war es, Möglichkeiten auszuloten, wie eine „verteilte digitale Bibliothek“ (wohl eigentlich eine Kombination aus virtueller Bibliothek und virtuellem Archiv) mit digitalisierten Zeugnissen der amerikanischen Geschichte geschaffen werden könne. Ein standardisiertes Format erleichtert auch die Archivierung und die Migration im Rahmen der digitalen Bestandserhaltung (MOA2 2001).

Ende des Jahres 2001 wurde das MOA2-Format dann von der Digital Library Federation[6] übernommen und zu METS, dem Metadata Encoding and Transmission Standard, weiterentwickelt (siehe Abschnitt 2.1.8).

[6] http://www.diglib.org/

2.1.8 Metadata Encoding and Transmission Standard (METS)

Der Metadata Encoding and Transmission Standard (METS) (METS 2003) entstand Ende 2001 aus dem MOA2-Standard (siehe Abschnitt 2.1.7) und berücksichtigte von Anfang an die im OAIS-Modell adressierten Verantwortlichkeiten für die digitale Langzeitarchivierung (OAIS 2002). Die Entwicklung des METS geht auf eine Initiative der Digital Library Federation zurück und wird von der Standardisierungsstelle der Library of Congress[7] betreut. Zahlreiche namhafte Bibliotheken und universitäre Institutionen in Nordamerika und Europa haben sich bereits als Anwender auf der Website des Standards registrieren lassen[8]. Es wird erwartet, daß der Standard, der derzeit bei der Version 1.3 angelangt ist, zunehmend breitere Anwendung finden wird (Ebind 2002). Von mehreren Bibliotheken sind Softwarewerkzeuge zur Verarbeitung von METS-Dokumenten angekündigt oder bereits verfügbar (z. B. METS-Java-Toolkit 2004).

Der METS kann als der generischste der bisher besprochenen Formate und Standards bezeichnet werden. Er ist ein Standard, um deskriptive, administrative und strukturelle Metadaten digitaler Bibliotheksobjekte in XML-Dateien zu kodieren. Er bietet dabei sehr weitreichende Möglichkeiten, die in dieser Arbeit nur soweit wie benötigt eingehender besprochen werden können. Für weitergehende Informationen verweise ich auf METS (2003). Ein Nachteil der zahlreichen Möglichkeiten sei auch nicht verheimlicht. Die Einarbeitung in das Format und seine Verwendung ist anspruchsvoll und benötigt Zeit.

Ein METS-Dokument besteht aus sieben Hauptabschnitten (METS 2003):

1. Der **Kopf** eines METS-Dokuments (Element `<metsHdr>`) enthält Metadaten, die das METS-Dokument selbst beschreiben (Meta-Metadaten), incl. Informationen wie Ersteller, Redakteur etc.

2. Ein Abschnitt mit **deskriptiven Metadaten** (Element `<dmdSec>`) kann im METS-Dokument eingebettete deskriptive Metadaten enthalten (z. B. Dublin-Core-, MAB- oder MABxml-Datensätze) oder auf deskriptive Metadaten verweisen, die außerhalb des METS-Dokuments vorliegen (z. B. ein MARC-Datensatz in einem OPAC) oder beides. Es können mehrere Datensätze externer oder interner deskriptiver Metadaten in diesem Abschnitt vorhanden sein.

3. Ein Abschnitt mit **administrativen Metadaten** (Element `<amdSec>`) enthält Informationen darüber, wie die primären Dateien erzeugt und gespeichert wurden, urheberrechtliche Angaben, Metadaten über das originale Quelldokument, von dem das Digitalisat stammt und Angaben zur Herkunft (*provenance*) der Dateien, die das gesamte digitale Dokument ausmachen (d. h. Beziehung von Master und Derivaten, sowie Migrations- und Transformationsinformationen; siehe Abschnitt 2.3, S. 14). Wie bei den deskriptiven Metadaten können diese Informationen außerhalb des METS-Dokuments oder darin eingebettet vorliegen.

4. Ein Abschnitt (Element `<fileSec>`) listet alle **Dateien** auf, die das digitale Dokument ausmachen. Dabei können mit `<fileGrp>`-Elementen Unterteilungen nach der Version vorgenommen werden (siehe Abschnitt 2.3.2, S. 16).

[7] `http://www.loc.gov/standards/`

[8] `http://sunsite.berkeley.edu/mets/registry/`

5. Ein Abschnitt mit **Strukturinformationen** (Element `<structMap>`) ist das Herz jedes METS-Dokuments. Er bezeichnet die hierarchische Struktur des digitalen Dokuments und verknüpft die Elemente dieser Struktur mit den Dateien, die den eigentlichen Inhalt ausmachen und mit Metadaten, die diese Elemente betreffen (siehe Abschnitt 2.3.1, S. 14).

6. Ein Abschnitt mit **strukturellen Verknüpfungen** (Element `<structLink>`) erlaubt es, das Vorhandensein von Hyperlinks zwischen einzelnen Knoten der hierachischen Strukturinformationen anzugeben. Dies ist besonders bedeutsam, wenn mit METS-Dokumenten Websites archiviert werden sollen.

7. Ein Abschnitt **Verhalten** (Element `<behaviorSec>`) kann verwendet werden, um Prozessierungsinformationen (ausführbares Verhalten) mit dem Inhalt des digitalen Dokuments zu verknüpfen. Jedes Verhalten (`<behavior>`) in diesem Abschnitt hat eine Schnittstellendefinition (`<interfaceDef>`), die eine abstrakte Definition der Prozessierung (d. h. des Verhaltens) darstellt und im Element `<mechanism>` einen Verweis auf ausführbaren Code, der die Prozessierung vornehmen kann. Die Prozessierungsinformationen können hierarchisch gegliedert werden. Einfache Beispiele für Prozessierungsinformationen wären etwa die Erstellung von PostScript-, DVI- (Devivce Independent) oder pdf-Dateien (Portable Document Format) aus einem TEX-Dokument für die Anzeige am Bildschirm oder den Ausdruck.

Das Format der METS-Dokumente wird durch eine XML-Schema-Definition festgelegt (METS Schema 1.3 2003). Dabei ist der Standard sehr offen für individuelle Anpassungen und zukünftige Bedürfnisse, da METS-Dokumente in nachvollziehbarer Weise um externe Metadatenformate erweitert werden können. Für eine Reihe von Angaben (z. B. bibliographische und technische) sind ausdrücklich Erweiterungen vorgesehen. Es wird empfohlen, statt proprietärer Eigenentwicklungen standardisierte Datenformate als Erweiterung zu nutzen. Die Einbindung standardisierter Erweiterungen vergrößert die Wahrscheinlichkeit, daß der komplette Inhalt eines METS-Dokuments längerfristig zugänglich und austauschbar bleibt. Als bevorzugte Erweiterungen des METS sind eine Reihe bereits bestehender, standardisierter Formate (z. B. Dublin Core[9], MARC[10], MARC XML[11], MODS[12]) aufgeführt (METS 2003; LoC 2004). Es wird von der METS-Gemeinde an der Entwicklung und Standardisierung weiterer, notwendiger Erweiterungsformate (z. B. NISO MIX) aktiv mitgearbeitet.

Die in den Metadaten beschriebenen, primären Dokumente können entweder in das METS-Dokument eingebettet werden oder aber als externe Dateien referenziert werden. Im ersten Fall bildet das METS-Dokument eine Einheit mit den primären Dateien, die als eine einzige, unter Umständen sehr große Datei gespeichert werden kann. METS-Dokumente eignen sich als Submission Information Package (SIP), Archival Information Package (AIP) und als Dissemination Information Package (DIP) im Sinne des OAIS-Modells (OAIS 2002).

Durch die Verwendung von XML (XML 2004) und XML-Schema (XML-Schema 2001) erscheint die Zugänglichkeit zum Inhalt der METS-Dokumente auf längere Zeit gesichert. METS-Dokumente können mit allen XML- und XML-Schema-verarbeitenden Standardsoft-

[9] http://www.dublincore.org/
[10] http://www.loc.gov/marc/
[11] http://www.loc.gov/standards/marcxml/
[12] http://www.loc.gov/standards/mods/

warewerkzeugen auf vielfältige Art und Weise verarbeitet werden. Ein Beispiel wird in Abschnitt 4.3 auf S. 35 gezeigt.

Die METS-Schema-Definition erlaubt sehr unterschiedlich strukturierte METS-Dokumente, was flexible Anpassungen an die jeweiligen Bedürfnisse erlaubt. Andererseits ist es für eine konkrete Software schwierig, alle Eventualitäten abzufangen, wenn in einem Anwendungsfeld eigentlich nur ein Ausschnitt all der Möglichkeiten benötigt wird. Deshalb wurde die Möglichkeit vorgesehen, sogenannte METS-Profile zu erstellen (McDonough 2003). In diesen können in Form einer XML-Datei strukturiert und detailliert Beschränkungen und Vorgaben beschrieben werden, denen eine bestimmte Menge von METS-Dokumenten unterliegt. Beispiele für derartige Vorgaben sind die Schemata verwendeter Erweiterungsformate, Formatvorgaben für die Dateien des digitalen Dokuments, kontrollierte Vokabularien und Beschreibungen der gestatteten Dokumentenstruktur. Solche METS-Profile sollen zukünftig bei der Library of Congress registriert und öffentlich gemacht werden können, um die Austauschbarkeit von METS-Dokumenten zu verbessern. In METS-Dokumenten besteht die Möglichkeit, das zugrundeliegende METS-Profil anzugeben.

2.2 Auswahl eines Metadatenstandards

Betrachtet man die oben angeführten Projekte zur Schaffung von Metadaten für digitale Bibliotheksobjekte, so ist die Tendenz erkennbar, daß versucht wird, von den anderen Projekten zu lernen und die Implikationen des OAIS-Funktionsmodells zu berücksichtigen. Bei den um Offenheit und breite Unterstützung bemühten Initiativen zeichnet sich eine Konvergenz hin zu zwei Standards ab, den OCLC/RLG Preservation Metadata und dem Metadata Encoding and Transmission Standard (METS). Von diesen beiden ist nur METS sofort einsatzbereit, da mit dem XML-Schema eine Grundlage für die programmatische Verarbeitung existiert. Für den METS stehen außerdem bereits frei verwendbare Softwarewerkzeuge bereit.

Nach Abwägung der Alternativen im Hinblick auf die in Abschnitt 1.3 dargestellte Zielsetzung komme ich zu dem Schluß, daß der Metadata Encoding and Transmission Standard (METS) die beste Grundlage für die Metadatenerhebung bei der Digitalisierung an der Univerversitätsbibliothek Regensburg ist.

Wesentliche Gründe für diese Entscheidung seien noch einmal kurz aufgeführt. METS-Dokumente eignen sich ausdrücklich als Submission Information Package (SIP), als Archival Information Package (AIP) und als Dissemination Information Package (DIP) im Rahmen des OAIS-Funktionsmodells. Der METS kann deshalb neben der Archivierung auch als universelles Export- und Importformat für digitale Dokumente dienen. Er bietet standardisierte (oder noch zu standardisierende) Möglichkeiten der Anpassung und Erweiterung, auch für noch nicht bekannte Dateiformate. METS-Dokumente eignen sich für die Beschreibung verschiedener Medien und Dokumententypen unterschiedlichster Granularität. Der METS besitzt eine breite Unterstützung durch Anwender und wird aktuell weiter gepflegt und entwickelt. Es ist bereits Software für die Erstellung und weitere Verarbeitung von METS-Dokumenten verfügbar und weitere ist absehbar. Nicht zuletzt können METS-Dokumente offline mit Dateien des primären Dokuments (z. B. auf CD-ROM) gespeichert werden.

Wie dargestellt, ist der METS vor allem ein Rahmenstandard, der um weitere Metadatenformate erweitert werden muß. Für bibliographische Metadaten werden das Maschinelle Austauschformat für Bibliotheken MAB (MAB2 1995) und Dublin Core (Dublin Core 2003)

zum Einsatz kommen (Details siehe Abschnitt 2.3.1, S. 14). Als technische Metadatenformate werden NISO MIX (MIX-Website 2003) für die bei der Digitalisierung vorherrschenden Einzelbilder und das Audio Technical Metadata Extension Schema (AUDIOMD 2002) der Library of Congress für Audioaufnahmen verwendet werden. Beide Formate haben sich u. a. bei der Library of Congress bewährt und sind ausdrücklich als Erweiterungsformate im METS-Schema aufgeführt. Der NISO MIX ist eine XML-Repräsentation (Cundiff 2003) der sich gerade in der Entwicklung befindlichen Norm NISO Z39.87 (NORM Z39.87-2002 2002).

Bei der weiteren Konzeption des Metadaten-Rahmenwerks für die Digitalisierung beziehe ich mich jetzt ausdrücklich auf den METS und die gewählten Erweiterungsformate und untersuche, wie die in Abschnitt 1.3 genannten Anforderungen damit gelöst werden können.

2.3 Sicherung der Authentizität von Dokumenten

Als Authentizität digitaler Dokumente werden je nach Kontext in sich stimmige, aber unterschiedliche Dinge bezeichnet (Cullen u. a. 2000):

- Die Eigenschaft, selber ein Original zu sein oder aber
- die Eigenschaft, ein analoges Original möglichst korrekt wiederzugeben,
- die Eigenschaft, von bekannter und nachvollziehbarer Herkunft zu sein oder
- die Eigenschaft, gegenüber einem Original unverändert (nicht manipuliert, verfälscht oder defekt) zu sein.

Der erste Aspekt ist hier nicht relevant, da nur Digitalisate, d. h. nicht genuin digitale Dokumente behandelt werden. Auch die zweite Auffassung von Authentizität ist nicht Gegenstand dieser Arbeit. Informationen darüber finden sich beispielsweise bei DLF & CLIR (2000) und Devlin u. a. (2004). Hier betrachte ich, wie die Abstammung des Digitalisats vom analogen Original beschrieben werden kann (Abschnitt 2.3.1), wie die Erzeugung digitaler Derivate der ersten digitalen Manifestation dokumentiert werden kann (Abschnitt 2.3.2) und wie sichergestellt werden soll, daß Kopien der ersten digitalen Manifestation einer Datei als unverändert erkannt werden können (Abschnitt 2.3.3). Den Begriff Authentizität verwende ich im folgenden nur in diesem letzten Sinne, der exakten (Bit für Bit) Übereinstimmung eines digitalen Dokuments mit seiner ersten Manifestation.

2.3.1 Abstammung vom analogen Original

Kernaufgabe von Metadaten für Digitalisate ist es, die Abstammung des digitalen Dokuments vom analogen Original zu kennzeichnen. Das bedeutet, das analoge Gesamtwerk überhaupt zu identifizieren, sowie zu dokumentieren, wie das Digitalisat gewonnen wurde und auch den Zusammenhang der bei der Digitalisierung oftmals zahlreich entstandenen Dateien untereinander und in ihrer Relation zum analogen Original zu erfassen. Die Abstammung verschiedener Dateien und Versionen des digitalen Dokuments voneinander (*digital provenance*) wird in Abschnitt 2.3.2 behandelt.

Die analoge Quelle für ein digitales Dokument kann im METS-Element <sourceMD> angegeben werden. <sourceMD> ist Teil der administrativen Metadaten und dafür konzipiert, Metadaten eines Erweiterungsformats entweder einzubetten oder als externe Quelle zu referenzieren. Das wichtigste bibliographische Metadatenformat im Rahmen dieser Arbeit ist MAB (MAB2 1995). An der Universitätsbibliothek Regensburg werden bisher fast ausschließlich in deutschen Bibliothekskatalogen verzeichnete Titel digitalisiert, für die MAB-Daten also bereits vorhanden sind. Ein MAB-Datensatz des Quelldokuments – wie im METS vorgeschrieben base64-kodiert – soll daher in einem <sourceMD>-Element die analoge Quelle angeben. MAB-Datensätze können beispielsweise aus dem OPAC des Bibliotheksverbundes Bayern kopiert werden. Zukünftig ist an eine Übernahme durch einen Z39.50-Client (z. B. JAFER, siehe JCDL'02 2002) gedacht. MAB ist zwar ein Binärformat, dessen Inhalt nicht direkt mit XML-verarbeitenden Werkzeugen erschlossen werden kann, aber es ist offen dokumentiert und wird sehr gut unterstützt. Seit kurzem existiert zudem unter dem Namen MABxml eine XML-Repräsentation von MAB (MABxml Website 2004). Sobald Softwarewerkzeuge für die Umwandlung von MAB in MABxml und zurück verfügbar sind, kann MABxml in die METS-Dokumente eingebunden werden. Die Informationen des MAB-Datensatzes können dann beispielsweise auch direkt für die Erstellung von Präsentationsformen aus dem METS-Dokument durch XSLT verwendet werden. Für analoge Quellen, für die keine MAB-Daten in Bibliothekskatalogen vorliegen, sollte ein Dublin-Core-Satz mit bibliographischen Angaben erzeugt und in einem <sourceMD>-Element gespeichert werden. Dublin Core (Dublin Core 2003) ist eines der ausdrücklich genannten Erweiterungsformate von METS.

Die Dokumentation, auf welche Art und Weise die Digitalisierung erfolgte, also mit welchen Geräten in welcher Einstellung, welche Software verwendet wurde etc., kann für spätere Nutzungen der Digitalisate von großer Bedeutung sein. Diese Angaben können bei METS-Dokumenten in technischen Metadaten gespeichert werden. Für Einzelbilder verwende ich in dieser Arbeit Metadaten im NISO-MIX-Format (Cundiff 2003), für Audiodaten das AUDIOMD-Schema (Audio Metadata Extension Schema) (AUDIOMD 2002). Diese beiden XML-basierten Formate bieten entsprechende Möglichkeiten und werden von der Library of Congress als Erweiterungsformate für METS-Dokumente verwendet (METS 2003; LoC 2004).

Die Abfolge der einzelnen Dateien des digitalen Dokuments gemäß ihrer Abstammung vom analogen Original kann mit Hilfe von Strukturmetadaten im METS-Element <structMap> festgelegt werden. So läßt sich beispielsweise die Seitenfolge einer digitalisierten Monographie unabhängig von Dateinamen eindeutig angeben. Die Möglichkeiten der Strukturmetadaten gehen aber weit darüber hinaus. Es können freie Hierarchien angegeben, mit Bezeichnungen und Sortierreihenfolge versehen, sowie jeweils mit deskriptiven und administrativen Metadaten und natürlich den eigentlichen Inhalten des Dokuments verknüpft werden. Sinnvoll ist, neben der Angabe der physischen Struktur (z. B. die Seitenfolge) auch die logische Struktur des Dokuments anzugeben. Diese kann beinahe beliebig tief erschlossen werden, wobei auch Teile von Dateien des Dokuments referenziert werden können. Bei logischen Hierarchien sind unterschiedliche Aspekte denkbar, die parallel nebeneinander erschlossen werden können. So könnte bei einer Monographie die Paginierung mit unterschiedlichen Seitenzählungen neben der Abfolge von Deckblatt, Schmutztitel, Titelseite, Vorwort, Inhaltsverzeichnis, Kapiteln etc. stehen. In Abschnitt 4.2, S. 33 ist ein Beispiel für die logische Erschließung eines Dokuments nach unterschiedlichen Aspekten gegeben. Diese Angaben sind hervorragend für die Erstellung von Präsentationsformen geeignet, die auf vielfältige Weise den bequemen Zugang zum Dokument ermöglichen (Beispiel siehe

Abschnitt 4.3, S. 35).

2.3.2 Digitale Derivate

Als Ergebnis der Digitalisierung eines analogen Objekts entsteht eine erste digitale Manifestation, die als Datei gespeichert wird. Dieser Datei, dem *digitalen Master*, kommt besondere Bedeutung zu, da sie die Eigenschaften des analogen Originals möglichst gut für alle beabsichtigten Nutzungen (im Sinne der Eignung für die *designated community* eines OAIS) widerspiegeln soll und Ausgangspunkt der Bemühungen der digitalen Bestandserhaltung ist. Entsprechend muß ein für die Medienform geeignetes Dateiformat gewählt werden, daß diesen Anforderungen bezüglich der Qualität und der Eignung für die Langzeitarchivierung genügt. Für Einzelbilder wird derzeit das Tagged Image File Format (TIFF 6.0 1992; NORM ISO 12234-2:2001 2001) präferiert, das dafür besonders geeignet erscheint (Frey 2000), für Audiodaten das WAVE-Format mit PCM-Daten (*Pulse Code Modulation*).

Vom digitalen Master können *digitale Derivate* gewonnen werden, die etwa im Rahmen der Langzeiterhaltung (Migration) notwendig werden oder für bestimmte Nutzungsarten geeigneter sind als der digitale Master. Im OAIS-Modell entspricht dies der systeminternen Pflege eines AIP bzw. der Erstellung eines DIP für den Zugang zur archivierten Information (OAIS 2002). Derivate werden z. B. Dateien in einem Präsentationsformat sein, deren Größe aus praktischen Gründen durch verlustbehaftete Komprimierung reduziert wurde oder kleine Vorschaubilder in einem Format, das mit möglichst vielen Anzeigeprogrammen (z. B. Web-Browser) dargestellt werden kann. Digitale Derivate können aber auch Dateien von einem anderen Medientyp sein, wie Texte mit Strukturauszeichnung, die durch OCR von Scans (Bilddateien) gewonnen wurden oder Abschriften (Text) von Sprachaufnahmen (Audio).

Die Informationen, welche Datei auf welche Weise vom digitalen Master – oder bereits von einem seiner Derivate – abgeleitet wurde, muß in den Metadaten gespeichert werden. Günstigenfalls sollte dies so erfolgen, daß Dateien, die, unabhängig vom Medientyp oder Format, die gleiche zugrundeliegende Information enthalten, leicht als Alternativen mit ihrer jeweiligen Nutzungsmöglichkeit erkannt werden können. METS-Dokumente bieten dazu mehrere Möglichkeiten, zum einen in der METS-Struktur selbst, zum anderen durch das Referenzieren oder Einbinden von Metadaten in anderen Formaten, die entsprechende Information tragen können.

In METS-Dokumenten kann jeder Datei ein `<digiprovMD>`-Element (*digital provenance metadata*) zugeordnet werden, in dem die Beziehungen von Master und Derivaten oder die Geschichte von Transformationen oder Migrationen der Datei vermerkt werden. Der Inhalt des `<digiprovMD>`-Elements und dessen Format sind im METS-Standard nicht vorgeschrieben. Es existiert meines Wissens nach bislang noch kein verbreiterter oder anerkannter Standard dazu. Der NISO-MIX-Standard bietet entsprechende Möglichkeiten für Bilddateien im Element `<ChangeHistory>` (Cundiff 2003). Deutlich erweiterte und nicht auf Bilddateien beschränkte Möglichkeiten wird der LMER-Standard (Long-term preservation Metadata for Electronic Resources) bieten, der bei Der Deutschen Biblothek entwickelt wird.

In einem METS-Dokument können alle Dateien, die einer bestimmten Version des Werks angehören in `<fileGrp>`-Elementen zusammengefaßt werden. Verschiedene `<fileGrp>`-Elemente können dann beispielsweise den digitalen Master, ein großformatiges Derivat im JPEG-Format, ein kleines Vorschaubild und eine Text-Version nach OCR enthalten. Anders herum können einander inhaltlich entsprechende Dateien in den verschiedenen Versionen

durch einen gemeinsamen Wert des Attributs `GROUPID` des `<file>`-Elements gekennzeichnet und damit leicht auffindbar gemacht werden. In Abschnitt 4.3 auf S. 35 wird gezeigt, wie dies zur Erstellung einer HTML-Seite mit Inhaltsangaben aus einem METS-Dokument durch XSL-Transformation genutzt wird.

2.3.3 Authentizität

Bei analogen Medien wird als Authentizität von Dokumenten – eigentlich der auf dem Medium gespeicherten Information – meist verstanden, daß es sich tatsächlich um das vermutete Dokument handelt und daß dessen Informationsgehalt nicht nachträglich verfälscht wurde. Diese Form der Authentizität wird zumeist dadurch als (mehr oder weniger gut) gesichert angesehen, daß sich Manipulationen an der enthaltenen Information als mehr oder weniger gut feststellbare Veränderungen am physischen Datenträger auswirken. Bei digitalen Medien wird als eine neuartige Bedrohung der Authentizität gesehen, daß sich Manipulationen sehr leicht spurenlos vornehmen ließen und damit unentdeckt bleiben könnten. Für eine Reihe analoger Medien, etwa Videos oder Tonkassetten, gilt dies natürlich genauso, was oft aber übersehen wird. Andererseits gilt dies nicht für endgültig schreibgeschützte digitale Datenträger, von denen etwa WORM-Datenträger (*Write Once Read Many*) speziell für diesen Zweck optimiert wurden (Gieselmann 2004). Als weiterer diesbezüglicher Vorbehalt ist wohl die Furcht vor „Hackern“ und „Crackern“ zu sehen, die von überall unkontrolliert in komplexe und deshalb vom Nicht-Spezialisten meist schlecht verstandene Informationssysteme eindringen und diese manipulieren könnten.

Digitale Signaturen dienen dem Nachweis der Herkunft eines digitalen Dokuments von der signierenden Person oder Institution und zugleich dem Beleg, daß das Dokument seit der Signierung unverändert ist. Digitale Signaturen haben aber keine dauerhafte Wirkung, sondern müssen regelmäßig erneuert werden, um mit dem raschen technologischen Wandel Schritt zu halten. Es bedarf deshalb eines ausgeklügelten Verwaltungssystems für signierte Dokumente, um die Signaturen gültig zu erhalten (Schulzki-Haddouti 2003a). Ein garantierter Herkunftsnachweis in dieser Form wird derzeit für die Digitalisate an der Universitätsbibliothek Regensburg nicht benötigt, so daß der Aufwand für den Einsatz digitaler Signaturen nicht gerechtfertigt erscheint.

Vielmehr muß eine Authentizitätsprüfung der hier betrachteten Digitalisate im wesentlichen nur der Fehlererkennung und ggf. -behebung dienen. So soll erkannt werden können, ob es sich bei einer Datei tatsächlich um die von den Metadaten referenzierte handelt oder ob sie unbeschädigt ist. In den in dieser Arbeit verwendeten Metadatenformaten (METS, NISO MIX) wird deshalb eine Prüfsumme der referenzierten Dateien gespeichert. Durch den Vergleich der gespeicherten mit einer neu berechneten Prüfsumme lassen sich Veränderungen an Dateien nachträglich erkennen und die Dateien können eindeutig identifiziert werden. Dies macht auch das Wiederfinden einer versehentlich umbenannten oder verschobenen Datei leicht möglich.

Eine Gewähr gegen Mißbrauch ist natürlich nur dann gegeben, wenn die ursprünglich berechnete Prüfsumme nicht genauso nachträglich verändert werden kann, wie das zu prüfende Dokument. Dies läßt sich am sichersten wohl durch zahlreiche redundante, schreibgeschützte Kopien der Prüfsumme an möglichst verschiedenen Orten erreichen, z. B. auch durch Papierausdrucke oder auf CD-ROMs. Dieses Prinzip redundanter, örtlich verteilter Kopien ist in Bibliotheken seit Jahrtausenden bewährt. Es gilt natürlich im Sinne der Sicherheit besonders auch für die eigentlichen Nutzdaten und kann dabei, quasi nebenbei, auch deren Authentizität sichern. Die Prüfsummen sind allerdings von sehr

viel kleinerem Umfang und somit einfacher zu handhaben. Im LOCKSS-Programm (*Lots Of Copies Keeps Stuff Save*) wird versucht das altbewärte Prinzip verteilter Kopien zur Datensicherung auch für digitale Dokumente zu etablieren (Reich 2002). Softwaretechnische Methoden zur Replikation digitaler Sammlungen mit dem Ziel der Bestandssicherung werden derzeit auch von Informatikern untersucht (z. B. Cooper und Garcia-Molina 2002). Die Verteilung von Kopien digitaler Dokumente und ihrer Prüfsummen läßt sich unter Umständen sehr viel einfacher und weniger fehleranfällig bewerkstelligen, als die Handhabung von technisch sehr aufwendigen, pflegebedüftigen und, wie die Vergangenheit immer wieder gezeigt hat, auch fehleranfälligen, elektronischen Sicherheitssystemen. Die Komplexität des Umgangs mit digitalen Signaturen wird als ein Haupthindernis für den bisher noch weitgehend ausgebliebenen wirtschaftlichen Erfolg dieser Technologie gesehen (Kossel 2003; Schulzki-Haddouti 2003b).

Die Anforderungen an die Authentizität der Digitalisate sind hier relativ gering, so müssen etwa keine gesonderten juristischen Belange beachtet werden. Es gibt keinen erkennbaren Grund für mißbräuchliche Datenveränderungen der Digitalisate. Ich erachte es deshalb als ausreichend für die Authentizitätswahrung, schreibgeschützte Kopien der Daten vorzuhalten, die offline und dem Zugriff über Datennetze entzogen sind. Realisiert wird dies dadurch, daß die Daten derzeit in mindestens zwei Kopien auf CD-ROM (fixiert, d. h. nicht wiederbeschreibbar) mit unterschiedlichem Lagerort gesichert werden (aber noch im selben Gebäude). Eine dieser CD-ROMs ist ein Archivexemplar, zu dem nur Mitarbeiter Zugang haben sollen. Wünschenswert ist es auch, Kopien an mindestens einem anderem Ort, z. B. bei der Bayerischen Staatsbibliothek zu lagern. Dies ist bisher erst bei einem Teil der Digitalisate der Fall.

Die Prüfsummen der Digitalisate (MD5-Verfahren) werden im Metadatensatz abgelegt und bei der Vergabe von Persistent Identifiern für Dateien in der PID-Datenbank hinterlegt (siehe Abschnitt 2.4, S. 18). Der Metadatensatz selbst wird derzeit nur zusammen mit den Digitalisaten gespeichert.

2.4 Persistente Identifier

Für eindeutige und dauerhaft gültige Bezeichner digitaler Objekte hat sich auch im Deutschen die Benennung *Persistent Identifier* bzw. Persistente Identifier (PID) eingebürgert (Schroeder 2003). In METS-Dokumenten werden für verschiedene Elemente eindeutige Bezeichner benötigt, damit innerhalb eines METS-Dokuments Verknüpfungen zu diesen Elementen geschaffen werden können (Beziehung des Typs `xs:ID` zu `xs:IDREF` bzw. `xs:IDREFS`; siehe XML-Schema 2001). Dies trifft vor allem auf `<file>`-Elemente zu, die eine Datei beinhalten oder referenzieren (bzw. auch mehrere identische Kopien einer Datei referenzieren können). Da es möglich sein muß, METS-Dokumente zu kombinieren, z. B. um auf getrennten CD-ROMs gespeicherte Teile eines umfangreichen Dokuments zusammenzuführen, sollten diese Bezeichner eindeutig über alle METS-Dokumente hinweg sein. Zusätzlich werden auch eindeutige Bezeichner für die METS-Dokumente selber benötigt. Also ist ein System zu schaffen, daß zentral für die Universitätsbibliothek Regensburg nach Bedarf solche Bezeichner generiert und und das programmatisch anzusprechen ist.

Die Digitalisate der Universitätsbibliothek Regensburg werden mangels eines entsprechend großen Serversystems derzeit auf CD-ROMs gespeichert. Da die Speicherkapazität von CD-ROMs bei großen Dokumenten oft nicht ausreicht, müssen die Dateien eines Dokuments ggf. auf mehrere CD-ROMs verteilt werden. Auf jeder CD-ROM wird deshalb ein

METS-Dokument mit dem Inhalt dieser CD-ROM gespeichert. Auf der ersten CD-ROM des Dokuments wird eine zusätzliche METS-Datei erzeugt, die alle METS-Dateien und damit das gesamte Dokument zusammenfaßt. Um auf den Speicherort eines anderen METS-Dokuments verweisen zu können, ist demnach aber die Kenntnis der CD-ROM notwendig, auf der es sich befindet. Es existiert in den Katalogsystemen des Bibliotheksverbundes Bayern kein eindeutiger und dauerhaft gültiger Bezeichner für Katalogeinträge oder für die katalogisierten Medien selbst (letzteres zumindest nicht an der Universitätsbibliothek Regensburg). Um die CD-ROMs dauerhaft eindeutig identifizieren zu können, muß deshalb ein eigenes System geschaffen werden, daß ihnen einen PID zuweist und das zusätzlich auf den Lagerort der CD-ROMs verweist. Dafür sind neue organisatorische und EDV-technische Einrichtungen nötig.

Der Standort mit dem Lokalkennzeichen 136 ist für CD-ROM-Sicherungsexemplare (und andere entspechende Medien) der Digitalisate reserviert. Als Neuerung im Geschäftsgang wird eingeführt werden, daß Medien von diesem Standort nur in besonderen Fällen umsigniert werden dürfen. Die Signaturen der CD-ROMs dieses Standorts sind deshalb geeignete Verweise auf den Lagerort. Sie werden zusammen mit der PID einer CD-ROM im PID-vergebenden System gespeichert. Mit Hilfe eines Abfragemechanismus im PID-System kann dann zu jedem PID einer CD-ROM der Lagerort in Erfahrung gebracht werden. Falls tatsächlich Umsignierungen von Medien dieses Standorts notwendig würden, müssen diese im PID-System mitgeführt werden. Eine Umstellung von den CD-ROMs als Speicherort auf ein Online-System ist problemlos möglich.

Von jeder neu erzeugten CD-ROM mit Digitalisaten wird eine Kopie als Archiv- und Sicherungsexemplar an diesem Sicherungsstandort 136 aufgestellt. Die Medien dieses Standorts sind als Sicherungsexemplare der Benutzung entzogen. Sie dienen jedoch als Kopiergrundlage, wenn bei der Benutzung beschädigte Medien ersetzt werden müssen. Es erscheint sinnvoll, die Medien dieses Standorts auch als Grundlage für die Bestrebungen der digitalen Langzeitarchivierung zu verwenden.

In den Metadaten muß dauerhaft zuverlässig auf die Dateien der digitalen Dokumente verwiesen werden. Da die Dokumente derzeit offline auf CD-ROM gespeichert werden, genügt die Angabe einer relativen URL innerhalb der CD-ROM, die vom METS-Dokument auf den Speicherort der Dateien verweist. Sobald die Dokumente zusammengeführt und mit den Metadaten auf ein Online-Speichersystem mit größerer Kapazität transferiert werden können, genügt diese Adressierung der Dateien vermutlich nicht mehr. Dann sollte der PID der Dateien für die Adressierung verwendet werden, d. h. als Schlüssel zum Speicherort. Dies ist leicht dadurch möglich, daß das System, das die PIDs vergibt so erweitert wird, daß es zu jedem PID auch den oder die Zugriffspfade zur Ressource speichert. Für jede PID können dann der oder die Speicherorte im Online-System ausgeliefert werden. Für die Dateien wird ja bisher schon eine Prüfsumme mit der vergebenen PID gespeichert (siehe Abschnitt 2.3.3, S. 17). Als zusätzliche Angabe kämen dann nur noch ein bis mehrere URLs dazu.

In der deutschen Bibliothekslandschaft setzen sich derzeit Uniform Resource Names (URN) (Moats 1997) unter Einbeziehung der National Bibliography Numbers (Hakala 2001) als standardisierte PIDs für digitale, online verfügbare Dokumente durch (Schroeder 2003; Hammen und Slotta 2004b, c, a). Sie haben, wie andere standardisierte PID-Systeme auch, den Vorteil, Ressourcen eine eindeutige, dauerhaft gültige Kennung zuzuweisen und diese vom tatsächlichen Speicherort zu trennnen. Zum Einsatz kommt dabei immer ein *Resolver*, d. h. ein Auflösungsmechanismus, der bei der Angabe des Persistent Identifiers entweder den tatsächlichen Speicherort der referenzierten Ressource (meist eine URL)

oder die Ressource selbst zurückliefert. Deshalb sollte auch das Resolvingsystem in Regensburg mit URNs arbeiten. Die Implementierung eines solchen URN-Resolvers erscheint unkompliziert.

Der Einsatz eines solchen Systems hat den weiteren Vorteil, daß von außen jede Datei zuverlässig und transparent referenziert werden kann. Die Digitalisate können deshalb auch ohne Probleme in andere Projekte integriert werden. So ist geplant, digitalisierte historische Pflanzenabbildungen (Digitalisierung des Sturm (1797–1862); siehe Abschnitt 1.1, S. 1) vom Informationsangebot des Botanischen Informationsknotens Bayern[13] aus einzubinden. Ebenso können Werke mit einem eigenen Eintrag im Bibliothekskatalog an Die Deutsche Bibliothek gemeldet werden, wo sie eine eigene URN erhalten. Als Angabe für den „lokalen Speicherort“ bei dieser Meldung bietet sich dann die Regensburger URN an, um Änderungen im Speicherort nur lokal pflegen und nicht immer an Die Deutsche Bibliothek melden zu müssen. Für die Einzeldateien der Dokumente erscheint eine URN-Meldung an Die Deutsche Bibliothek wegen zu geringer Granularität nicht sinnvoll; der lokale Resolver genügt hier.

[13] http://www.bayernflora.de/

3

Prototypische Implementierung

Bisher waren keine Programme verfügbar, mit denen die im vorhergehenden Abschnitt konzipierten Aufgaben in für die Universitätsbibliothek Regensburg geeigneter Weise erledigt werden konnten. Deshalb programmierte ich einen Grundstock von Softwarewerkzeugen, mit denen METS-Dokumente erstellt, Persistente Identifier vergeben und Präsentationsformen aus den METS-Dokumenten erzeugt werden können.

Desweiteren erforderte die neuartige Erschließung organisatorische Veränderungen bei der Digitalisierung, die in diesem Abschnitt zusammen mit den Hardwareanforderungen für das entwickelte System dargestellt werden.

3.1 Programmierung eines Erschließungswerkzeugs

Bisher existierten keine fertigen Programme für die Erschließung von Digitalisaten in der Form von METS-Dokumenten. Vielmehr gab es nur Module, die als Komponenten in einer selbst zu erstellenden Software nicht mehr als die Verarbeitung von METS-Dokumenten ermöglichen. Wie dabei die eigentliche Verarbeitung der Daten erfolgt, woher diese Daten stammen oder wie sie neu generiert werden, mußte selbst gelöst werden. Ich mußte also die Erzeugung und Speicherung von METS-Dokumenten sowie die Erfassung von bibliographischen Metadaten (MAB, Dublin Core), Strukturmetadaten und technischen Metadaten programmieren. Da die Erfassung von Metadaten bei der Digitalisierung den Digitalisierungsprozeß zumeist unzumutbar unterbrechen und verzögern würde, sollte die Erstellung der METS-Dokumente erst beginnen, wenn nach der Digitalisierung bereits alle Dateien vorliegen.

Die METS-Dokumente sollen zusammen mit den Digitalisaten dezentral auf CD-ROMs gespeichert werden (siehe Abschnitt 3.3, S. 27). Deshalb kam primär kein datenbankgestütztes System zur Erfassung der Metadaten in Frage. Die METS-Dokumente können aber jederzeit nachträglich in ein entsprechendes Datenbanksystem importiert werden. Das Erschließungswerkzeug speichert derzeit METS-Dokumente im Dateisystem. Erweiterungen des Programms, um METS-Dokumente über das Netzwerk an einen Server zu übertragen, dabei Webservices zu nutzen etc. sind jederzeit möglich.

Neben den Programmteilen für die automatisch ablaufenden Vorgänge sollte auch noch eine graphische Benutzeroberfläche geschaffen werden. Ich mußte die Wahl treffen zwischen einer Webanwendung mit dynamischen HTML-Seiten, bei der die eigentliche Arbeit des Programms auf dem Server abläuft und einem dezentralen Programm mit graphischer Benutzeroberfläche, das auf jedem Client-Rechner laufen kann. Ich entschied mich für die

zweite Variante, da das Programm bei dieser sehr viel flexibler einsetzbar ist. Viele Funktionen sind an einem Client-Rechner einfacher zu implementieren, da dort die direkte Interaktion mit dem Programm, unmittelbare Konsistenzprüfungen (z. B. von Benutzereingaben) und Schnittstellen zur Scansoftware leichter realisierbar sind.

Als Programmiersprache wählte ich Java[1]. Java-Programme sind weitgehend plattformunabhängig, das heißt auf verschiedenen Betriebssystemen und Hardwarearchitekturen einzusetzen. Aufgrund objektorientierter Programmierung läßt sich die geschaffene Software modular verwenden, was hier von großer Wichtigkeit ist. In Java lassen sich zudem gut graphische Oberflächen programmieren. Die XML-Verarbeitung ist mit Java sehr erleichtert, da diesbezüglich eine reichhaltige Funktionalität zur Verfügung steht. Mit dem Castor-Framework[2] ist beispielsweise die automatische Generierung von Java-Klassen aus einem XML-Schema heraus möglich, mit denen bequem dem XML-Schema entsprechende XML-Dokumente verarbeitet werden können. Ich entschied mich allerdings für den Einsatz des bereits programmierten METS Java Toolkit (METS-Java-Toolkit 2004) von Stephen Abrams (Harvard University Library). Zur meiner Entscheidung für Java als Programmiersprache trugen natürlich auch meine Erfahrungen und Kenntnisse bei.

Die Software ist komplett modular gestaltet, um jederzeit Erweiterungen oder Veränderungen zu erlauben. Realisiert ist das unter anderem durch Interfaces, die stellvertretend für konkrete Klassen stehen, die die Funktionalität auf unterschiedliche Weise implementieren können. So gibt es verschiedenen Parser, die aus den Dateinamen des digitalen Dokuments Strukturmetadaten extrahieren. Durch die Wahl eines geeigneten Parsers für Dateinamen, können diese sehr flexibel und angepaßt an das jeweilige Objekt vergeben werden. Die Gewinnung technischer Metadaten aus den Dateien hängt natürlich vom Dateiformat ab. Deshalb wird abhängig vom MIME-Type (Freed und Borenstein 1996) einer Datei eine Klasse gesucht, die dies für die gegebene Datei leisten kann.

3.1.1 Ablauf der Erschließung

Die Erschließung soll zuerst so weit wie möglich automatisch ablaufen. Daran können sich dann eine intellektuelle Prüfung des Ergebnisses und ggf. eine weitere intellektuelle Erschließung anschließen.

Die grundsätzlichen Schritte der automatischen Erschließung:

1. Dateien auflisten und im Element `<fileSec>` eintragen; dabei PIDs holen.

2. Technische Metadaten der Dateien extrahieren; die Metadaten werden im `<fileSec>`- und in `<techMD>`-Elementen gespeichert.

3. Strukturmetadaten erstellen: Dateinamen parsen oder aus Angaben aus einer externen Quelle übernehmen.

4. Ergebnisse zusammenführen in einem METS-Dokument; dabei METS-PID holen.

Die Schritte der anschließenden intellektuellen Erschließung:

1. Bibliographische Angaben hinzufügen, entweder durch Neuanlegen, meist aber durch Übernahme vorhandener Metadaten.

[1] `http://java.sun.com/`
[2] `http://castor.exolab.org/`

2. Ergebnisse überprüfen und ggf. korrigieren.

3. Ggf. weitere, tiefere strukturelle Erschließung.

3.1.2 Dateien, technische und strukturelle Metadaten

Die Auflistung und Erfassung aller Dateien des digitalen Dokuments im Element <fileSec>, die Gewinnung technischer Metadaten und die erstmalige Erzeugung struktureller Metadaten aus den Dateinamen hängen direkt voneinander ab. Deshalb sind diese Aufgaben am effizientesten gemeinsam zu erledigen.

Bei allen drei Aufgaben werden die Dateien, bzw. die neugeschaffenen Verweise im <fileSec>-Element auf diese und die Dateinamen benötigt. Schon bei der Auflistung der Dateien werden einige technische Metadaten (MIME-Type, Prüfsumme, Dateigröße, Dateidatum) im <fileSec>-Element gespeichert. Beim Holen von PIDs muß jeweils die Prüfsumme der Datei mit angegeben werden.

Der Ablauf sieht folgendermaßen aus:

1. Auflisten aller Dateien durch ein eigenes Modul. Das Auflisten erfolgt derzeit durch ein einfaches Verzeichnislisting, kann nach Bedarf aber auch anders gelöst werden, z. B. aus einer extern erzeugten, strukturierten Textdatei wie bei Ebind (Ebind 2002). Die Reihenfolge der Dateien in der erzeugten Liste muß die natürliche, d. h. dem Dokument entsprechende, Abfolge der Dateien sein, denn diese lineare Abfolge wird in die Strukturmetadaten übernommen.

2. Für jede Datei:

   (a) Extraktion minimaler technischer Metadaten (MIME-Type, Prüfsumme, Dateigröße, Dateidatum), Holen einer PID mit Speicherung der Prüfsumme im PID-System, Speicherung der Daten in einem <file>-Element.

   (b) Anhand des MIME-Types finden einer Klasse, die technische Metadaten für Dateien dieses MIME-Types extrahiert. Extraktion technischer Metadaten aus der Datei und Extraktion technischer Metadaten zum Digitalisierungsprozeß aus einer Konfigurationsdatei. Es werden also Informationen aus der Datei und aus einer für das zur Digitalisierung verwendete Gerät spezifischen Konfigurationsdatei zusammengeführt. Speichern der Metadaten in einem MIME-Type-spezifischen Format in einem <techMD>-Element.

   (c) Extraktion von Strukturmetadaten aus dem Dateinamen. Zum Parsen der Dateinamen existieren mehrere Module, die objektspezifisch für ein bestimmtes Muster des Dateinamens Informationen entnehmen können. Dadurch ist angepaßt an das digitale Dokument die einfache Generierung von Strukturmetadaten möglich. Die Erzeugung von Strukturmetadaten aus geeignet gewählten Dateinamen hat sich als sehr effektiv und effizient erwiesen und soll deshalb unbedingt beibehalten werden. In vielen Fällen ist man bei der Scansoftware auf proprietäre Programme der Scannerhersteller ohne definierte Schnittstellen angewiesen. Der Transport von Strukturinformationen über die Dateinamen ist häufig der einzige Weg für die teilweise Automatisierung. Auf alle Fälle wird immer die lineare Abfolge der Dateien auch in den Strukturmetadaten erfaßt.

3. Nach Bedarf weitere Extraktion von Strukturmetadaten aus einer externen Quelle (Text- oder XML-Datei, Datenbank; bisher nicht implementiert).

4. Zusammenführen der verschiedenen Elemente im METS-Dokument.

3.1.3 Bibliographische Metadaten

Bibliographische Metadaten werden sowohl für die analoge Quelle des Digitalisats (im `<sourceMD>`-Element) als auch für das digitale Dokument selbst (im `<dmdSec>`-Element) erfaßt. Das Programm wird verschiedene Module anbieten, die jeweils ein Metadatenformat aus jeweils einer Quelle liefern können.

So wird ein Modul es ermöglichen, MAB-Datensätze durch Kopieren aus dem Verbundkatalog des BVB zu holen. Dieser OPAC, der auch die Regensburger Titel nachweist, bietet die Ausgabe von Titeldatensätzen in diesem Format an. Ein anderes Modul soll MAB-Daten später über das Z39.50-Protokoll[3] direkt holen. Als Grundlage für die Implementierung des Z39.50-Clients wird wohl das JAFER-Toolkit dienen (JCDL'02 2002).

Falls statt einer vollen Titelaufnahme in MAB nur Daten im Dublin-Core-Format gespeichert werden sollen, so wird ein Modul ermöglichen, diese in einem Formular der graphischen Benutzeroberfläche manuell zu erfassen. Bei größeren Digitalisierungsprojekten werden spezifische Metadaten oft in einer eigenen Datenbank erfaßt. In solchen Fällen wird sich die Programmierung eines eigenen Moduls lohnen, das Dublin-Core-Metadaten aus der Datenbank des Projekts holt.

3.1.4 Graphische Benutzeroberfläche

Die im Rahmen dieser Arbeit enwickelten Programme sind bisher nur von der Kommandozeile aus zu bedienen. Eine graphische Benutzeroberfläche ist noch nicht implementiert. Die kommandozeilenorientierten Programme haben den Vorteil, in Batch-Jobs eingesetzt werden zu können, mit denen sehr effizient große Mengen von Digitalisaten, insbesondere auch schon vorhandene, mit Metadaten erschlossen werden können. Mit den funktionierenden Klassen ist außerdem die Grundlage für die Programmierung der graphischen Benutzeroberfläche mit einer mehrschichtigen Architektur nach dem *Model-View-Controller*-Prinzip gelegt. Die existierenden Klassen stellen dabei die Model-Schicht dar. Diese Form der Mehrschichtarchitektur entspricht aktuellen Empfehlungen für die Softwareentwicklung, denn dadurch sind Erweiterungen und Veränderungen der Software am leichtesten zu leisten.

Primäre Zielsetzung der graphische Oberfläche ist, die Kontrolle und Bearbeitung der unübersichtlichen XML-Dokumente zu erleichtern, in dem viel von der Komplexität der Syntax versteckt wird und die wesentlichen Aspekte hervorgehoben werden. Einige Aspekte der Darstellung und Bearbeitung von METS-Dokumenten testete ich als *proof of concept* bereits aus. Die Darstellung der hierarchische Struktur der Strukturmetadaten in einer Baumansicht bewährte sich dabei sehr, denn so bleibt die komplexe Struktur übersichtlich und es kann gut darin navigiert werden. Auch das Editieren der Strukturmetadaten (Hinzufügen, Verschieben und Löschen von Elementen oder Teilbäumen) ist so einfach lösbar. Java bietet im Swing-Toolkit graphische Bedienelemente, die visuelle Entsprechungen eines externen Datenmodells darstellen. Manipulationen an den graphischen Elementen können

[3] `http://www.loc.gov/z3950/agency/`

so ihren direkten Niederschlag in den Daten finden und auch umgekehrt. Dieses Konzept erleichtert die graphische Programmierung sehr.

Bei der Erschließung in der graphischen Oberfläche sollte natürlich auch das digitale Dokument selbst angezeigt werden können. Die Funktionen für die Anzeige werden modular gestaltet werden, um mit einer Art Plug-Ins die Anzeige und Erschließung von Dokumenten verschiedener Medientypen und Dateiformate zu gestatten.

3.1.5 Erzeugung von Präsentationsformen

Durch die Verwendung von XML-kodierten Metadaten stehen diese auf transparente und, angesichts der komplexen Struktur der METS-Dokumente, auch auf relativ problemlose Weise für die Weiterverarbeitung zur Verfügung. Als einfaches Beispiel soll hier die Erzeugung eines HTML-Inhaltsverzeichnisses einer CD-ROM mit Hilfe der Technik XSLT (XSLT 1999) dienen. Dazu schrieb ich ein XSL-Stylesheet, das die Anweisungen für die Transformation eines METS-Dokuments in eine HTML-Datei enthält. Über XPath-Anweisungen (XPath 1999) kann in einem solchen Stylesheet jeder Bestandteil einer XML-Datei referenziert und an beliebiger Stelle in einem Ausgabedokument ausgegeben werden. Darüberhinaus sind auch weitere Operationen, wie Sortierungen, Filter, Berechnungen etc. möglich. Das Ergebnis einer solchen Transformation eines METS-Dokuments in eine HTML-Datei ist in Abschnitt 4.3, S. 35 dargestellt.

Dieses Beispiel hat einen großen praktischen Nutzen, können doch mit einem derartigen Inhaltsverzeichnis den Benutzern die Inhalte der CD-ROMs mit den Digitalisaten bequem zugänglich und transparent gemacht werden. Vor allem der Zugang über die Struktur der digitalen Dokumente bietet eine deutliche Verbesserung gegenüber dem bisherigen Zustand und kann unter Umständen besser sein, als die bisher schon erstellten pdf-Dateien mit ihrer rein linearen Abfolge der Seiten ohne weitere Strukturinformation.

pdf-Dateien eignen sich allerdings sehr gut zum Ausdrucken von mehrseitigen Dokumenten. Sie werden bisher mit dem Programm Adobe Acrobat erstellt. pdf-Dateien können aber auch mit XSLT über *XSL Formatting Objects* aus XML-Dateien erstellt werden. Es ist noch zu prüfen, welche Graphikformate dabei verarbeitet werden können und ob die Strukturinformationen in den METS-Dokumenten zur Erstellung von pdf-Bookmarks verwendet werden können, um mit diesen die Navigation in den pdf-Dateien erheblich zu verbessern. Dann gäbe es eine weitere Möglichkeit, eine benutzerfreundliche Präsentationsform der digitalen Dokumente direkt aus den METS-Dokumenten zu generieren.

3.2 Programmierung eines Systems für Persistente Identifier

Zur zentralen Vergabe und Verwaltung von Persistent Identifiern für Dateien, METS-Dokumenten und CD-ROMs implementierte ich ein datenbankgestütztes System mit HTTP-Schnittstelle. Als Programmiersprache kam PHP[4] zum Einsatz, als Datenbank MySQL[5]. Das PHP-Skript läuft auf einem Apache-Webserver[6]. Sobald das System gründlich ausgetestet ist, wird es produktiv eingesetzt und ab diesem Zeitpunkt zentral für die Univer-

[4] http://www.php.net/
[5] http://www.mysql.com/
[6] http://www.apache.org/

sitätsbibliothek Regensburg PIDs vergeben und verwalten. Die Erweiterung um Resolving-Mechanismen bis hin zu einem URN-Resolver ist problemlos möglich. Der Zugang zum PID-System wird IP-Adressen-basiert und durch Paßwortschutz kontrolliert werden.

Ein Client kann mit einfachen GET- oder POST-Requests nach dem HTTP-Protokoll (Fielding u. a. 1999) bei dem System PIDs für Ressourcen anfordern. Je nach Ressource, für die ein PID gewünscht wird (Dateien, METS-Dokumente oder CD-ROMs), müssen weitere, obligatorische oder können fakultative Parameter beim Request mit angegeben werden. Ist der Request korrekt, werden ein HTTP-Status-Code „`200 OK`“ und im Body der HTTP-Response der gewünschte PID zurückgeliefert. Bei fehlerhaften Anfragen werden ein HTTP-Status-Code „`400 Bad Request`“ und eine kurze Fehlermeldung im Body zurückgeliefert. Falls interne Probleme auftauchen (z. B. Datenbankfehler), wird mit einem HTTP-Status-Code „`500 Internal Server Error`“ geantwortet. Jeder Client, der diese einfachen HTTP-Befehle verarbeiten kann, kann PIDs anfordern. Dies ist auch von Hand mit einem Web-Browser möglich. Das hier implementierte Erschließungswerkzeug verfügt bereits über eine entsprechende Funktionalität zum Holen von PIDs.

3.2.1 Persistent Identifier für CD-ROMs

Die Vergabe von PIDs für CD-ROMs ist der komplexeste Fall, da CD-ROMs Projekten zugeordnet werden und dies im PID in Form eines Präfix kodiert wird. Die Organisation der CD-ROMs nach Projekten entspricht praktischen Überlegungen und den Wünschen der beteiligten Mitarbeiter. Projekte können zusammen mit einer kurzen Beschreibung im PID-System angelegt werden. Bei der Anforderung eines PID muß der Präfix des Projekts mit angegeben werden. Die Angabe der Signatur der CD-ROM ist optional. Der PID einer CD-ROM entspricht dem Muster „`PRÄFIX_123456`“. Innerhalb eines Projekts werden die PIDs durch Hochzählen vergeben, wobei die Angabe der Nummer immer sechsstellig ist.

Für die bequemere Verwaltung von CD-ROMs, ihren Projekten und Signaturen mit HTML-Formularen im Web-Browser wird ein zusätzliches PHP-Skript entwickelt.

Die Syntax für eine beispielhafte HTTP-Anfrage lautet:

```
GET /pfad/pid.php?action=generate_new_cd_id&cd_prefix=MMZ-2004\
&sig=136%2FWL+1080+W424-1%2C1 HTTP/1.0
```

3.2.2 Persistent Identifier für METS-Dokumente

Die PIDs für METS-Dokumente werden vom System als positive Ganzzahlen ausgeliefert, die im System einfach hochgezählt werden. Bei ihrer Verwendung als ID-Attribut in den METS-Dokumenten wird die Zeichenkette „`METS`“ davorgehängt. So werden sie eindeutig als METS-PIDs gekennzeichnet. Außerdem dürfen XML-Attribute vom Typ „`xs:ID`“ nicht mit einer Zahl beginnen (XML 2004).

Bei der Anfrage nach einem neuen METS-PID sollte die PID der CD-ROM, auf der das Dokument gespeichert wird, angegeben werden. Im Falle hierarchisch gegliederter METS-Dokumente ist die Angabe des PID des Vaterdokuments möglich. Zusätzlich kann ein Kurztitel des METS-Dokuments mitgegeben werden. Diese Zusatzinformationen werden in der Datenbank des PID-Systems gespeichert und sollen die Abfrage und Verwaltung erleichtern. Eine Erweiterung der mit der METS-PID gespeicherten Informationen um Speicherorte als Basis eines URN-Systems ist jederzeit einfach möglich.

Die Syntax für eine beispielhafte HTTP-Anfrage lautet:

```
GET /pfad/pid.php?action=generate_new_mets_id&cd=MMZ-2004_000123\
&father=123&short_title=Flora+von+Bayern%2C+Band+11 HTTP/1.0
```

3.2.3 Persistent Identifier für Dateien

Analog zu den METS-PIDs werden die PIDs für Dateien als posititve Ganzzahlen ausgeliefert, vor die in METS-Dokumenten die Zeichenkette „`FILE`“ gesetzt wird.

Bei der Anforderung eines neuen Datei-PID müssen die MD5-Prüfsumme der Datei und der PID des zugehörigen METS-Dokuments mit angegeben werden. Die Prüfsumme dient unter anderem der Authentizitätssicherung der Dateien (siehe Abschnitt 2.3.3, S. 17).

Die Syntax für eine beispielhafte HTTP-Anfrage lautet:

```
GET /pfad/pid.php?action=generate_new_file_id&mets=123\
&md5=b6e7503ed638d2471f9efb0e2fb320ce HTTP/1.0
```

3.3 Organisatorische Veränderungen bei der Digitalisierung

Die neu konzipierte Erschließung muß auch in den Geschäftsgang der Digitalisierung passen und darin integriert werden. Ich werde zuerst den bisherigen Ablauf der Digitalisierung skizzieren, um danach darzustellen, welche Neuerungen hinzukommen.

3.3.1 Bisheriger Geschäftsgang

Es lassen sich die folgenden wesentlichen Schritte unterscheiden:

1. Digitalisierungsauftrag geht von einem Benutzer oder Mitarbeiter ein.
2. Entscheidung über das geeignete Gerät (ggf. auch mehrere) für die Aufgabe.
3. Festlegung, wer führt die Digitalisierung wann durch.
4. Festlegung des Dateinamensschemas um ggf. Dokumentenstruktur zu kodieren.
5. Digitalisierung, ggf. durch Hilfskräfte.
6. Erzeugte Dateien werden ggf. über das Netzwerk auf einen anderen Rechner für die Weiterverarbeitung verschoben.
7. Eine leitende Mitarbeiterin überprüft stichprobenartig die Scanqualität und die Kollationierung, d. h. ob alle Seiten gescannt und korrekt bezeichnet sind.
8. Ggf. Erzeugen von Metadaten in einer Textdatei: Bibliographische Angaben, verwendeter Scanner, Dateiformat(e), ggf. Farbprofil.
9. Ggf. Erzeugen einer pdf-Datei als Präsentationsformat.
10. Sichern der Daten auf CD-ROM.

3.3.2 Neuerungen

Der Mehraufwand für die Erschließung ist gegenüber dem bisherigen Ablauf aufgrund der weitgehenden Automatisierung nur gering und die Neuerungen passen sich gut in den Geschäftsgang ein. Wie bisher wird das Dokument mit den Metadaten erst nach Abschluß der Digitalisierung erstellt. Die Kodierung der Struktur der Dokumente (bzw. von Teilen davon) in den Dateinamen erfolgt aber, auch wie bisher schon, bereits bei der Digitalisierung.

1. Die Benennung der Dateinamen muß sich an vorhandenen Möglichkeiten zur Extraktion von Strukturmetadaten, d. h. an den bereits programmierten Parsern, orientieren. Dies stellt aber keine wirkliche Einschränkung dar. Es ist weiterhin möglich, die Benennung der Dateien flexibel in Anpassung an die Objekte zu handhaben. Mit gewissen Grundmustern der Benennung kann in den Dateinamen kodierte Information ausgesprochen einfach zur Erzeugung von Strukturmetadaten genutzt werden. Was kodiert wird, soll am jeweiligen Objekt entschieden werden (z. B. gewünschte Erschließungstiefe, relativer Aufwand für individuelle Vergabe von Dateinamen im Gegensatz zu einem einfachen Hochzählen eines Nummernteils der Dateinamen durch die Scansoftware).

2. Die Handhabung der Digitalisate nach dem Fertigstellen der Dateien wird umorganisiert. Jeder Digitalisierungsplatz erhält ein eigens für ihn mit Samba[7] freigegebenes Verzeichnis auf einem Linux-Server. Der Digitalisierer verschiebt die Dateien nach der abgeschlossenen Digitalisierung von der lokalen Festplatte des Digitalisierungsrechners auf dieses Netzlaufwerk. Eine besonders geschulte Mitarbeiterin kann von ihrem Arbeitsplatz aus auf alle diese Verzeichnisse auf dem Dateiserver zugreifen. Sie verschiebt die Dateien in ihren eigenen, für die Digitalisierungsplätze nicht zugänglichen Arbeitsbereich auf dem Dateiserver und kopiert eine vorbereitete XML-Datei mit gerätespezifischen technischen Metadaten dazu. Dadurch kennzeichnet sie die Herkunft der Dateien vom jeweiligen Scanner. Die XML-Datei mit den gerätespezifischen Metadaten wird vom Erschließungsprogramm eingelesen und der Erzeugung der technischen Metadaten verwendet. Anschließend prüft die Mitarbeiterin die Qualität und die Kollationierung der Digitalisate wie bisher schon, erstellt die Metadaten und sichert die Dateien auf CD-ROM.

   Der Vorteil dieses Vorgehens ist neben der sofortigen Zuordnung der richtigen gerätespezifischen Metadaten zu den Dateien auch die Zwischenspeicherung der umfangreichen Digitalisate. Durch das Verschieben weg vom Digitalisierungsarbeitsplatz ist leichter zu organisieren, welche Dateien fertig für die weitere Verarbeitung sind. Es sammeln sich weniger Kopien der Digitalisate, die unter erheblichem Koordinierungsaufwand der Mitarbeiter in überflüssig oder noch benötigt geschieden werden müssen.

3. Die Vergabe der Metadaten liegt in den Händen einer besonders geschulten Mitarbeiterin. Falls benötigt teilt sie die Dateien des Digitalisats in Gruppen auf, die auf CD-ROMS passen und besorgt Persistent Identifier für die zu erstellenden CD-ROMs. Für jede CD-ROM wird nun mit dem Erschließungsprogramm eine METS-Datei erstellt, wobei automatisch Strukturmetadaten, technische und bibliographische Metadaten erfaßt werden. Nach dem automatischem Programmablauf sollen die METS-Dokumente zukünftig in der graphischen Benutzeroberfläche weiter bearbeitet, d. h.

[7] http://www.samba.org/

tiefer erschlossen oder korrigiert werden können. Anschließend werden die METS-Dokumente bei den Digitalisaten gespeichert und zusammen mit diesen auf CD-ROM gesichert.

4. Als eine zusätzliche Möglichkeit kann vor dem Brennen einer CD-ROM aus dem METS-Dokument heraus noch ein Inhaltsverzeichnis der CD-ROM erstellt werden (siehe Abschnitt 4.3, S. 35).

3.4 Hardwareanforderungen

Abschließend sei noch kurz auf die Hardwareanforderungen für den Einsatz der einzelnen Softwarekomponenten eingegangen.

Der Server für das PID-vergebende System sollte ständig erreichbar sein, um keine Verzögerungen bei der Arbeit zu verursachen. Die Last auf dem Server durch das PID-System ist vernachlässigenswert, so daß es auf dem Webserver oder einem der anderen Server der Bibliothek mitbetrieben werden wird. Es eignet sich grundsätzlich jeder Webserver mit PHP, MySQL und bevorzugt Apache als HTTP-Server.

Zum Zwischenspeichern der Digitalisate bei der Bearbeitung reicht ein kleiner Linux-Server (ab Pentium) mit dem Server-Programm Samba. Zum Zwecke der Datensicherheit sollte der Dateiserver zumindest mit einem Software-RAID und zwei Festplatten zur Datenspiegelung betrieben werden. Es genügt sicherlich der Einsatz von zwei aktuellen Desktop-Festplatten, da diese ausreichend Platz bieten und zuverlässig genug sind.

Für die Erschließung selber eignet sich jeder nicht zu alte Personalcomputer mit Windows oder Linux als Betriebssystem. Ein Anschluß an das Datennetz ist bei diesen Rechnern Voraussetzung, um Persistent Identifier vom Server zu holen und um mit dem Samba-Server arbeiten zu können. Dies ist an der Universitätsbibliothek Regensburg aber ohnehin der Fall. Bei Verwendung der in Java geschriebenen graphischen Oberfläche sollte der Arbeitsspeicher nicht zu knapp bemessen sein, um ein flüssiges Arbeiten zu erlauben.

4

Exemplarische Erschließung von Digitalisaten

In diesem Abschnitt möchte ich kurz die Ergebnisse der Erschließung in METS-Dokumenten mit den geschaffenen Programmen demonstrieren.

4.1 Komplette Erschließung eines Einzelbildes

Zuerst stelle ich das komplette Ergebnis der Erschließung des Digitalisats einer historischen Landkarte dar. Bei der Digitalisierung am Großformatscanner entstand nur eine einzige Datei. Das Digitalisat wurde mit bibliographischen Angaben, der Bezeichnung der analogen Quelle, technischen Angaben zum Digitalisierungsprozeß, technischen Angaben zum digitalen Master und mit Strukturmetadaten erschlossen. Letztere fallen bei einer einzelnen Datei naturgemäß nur minimal aus.

In der folgenden METS-Datei ließ ich zur besseren Übersicht die base64-kodierten MAB-Daten weg und kürzte lange Wertelisten in den technischen Metadaten.

```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<mets
  ID="METS42"
  LABEL="Topographischer Atlas vom Königreiche Baiern. Band [2]: Voelkershausen.
   Reducirt und gezeichnet von Schwarzmann. Berge gestochen von Georg Mayr.
   [ca. 1830-1840]"
  xmlns="http://www.loc.gov/METS/"
  xmlns:xlink="http://www.w3.org/TR/xlink"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.loc.gov/METS/
    http://www.loc.gov/standards/mets/mets.xsd">
  <metsHdr CREATEDATE="2004-05-17T10:14:23">
    <agent ROLE="CREATOR" TYPE="ORGANIZATION">
      <name>Universitätsbibliothek Regensburg</name>
    </agent>
  </metsHdr>
  <dmdSec ID="METS42-dmdSec1" CREATED="2004-05-17T10:14:03">
    <mdWrap MDTYPE="OTHER" OTHERMDTYPE="MAB2"
      LABEL="Datensatz im Maschinellen Austauschformat für Bibliotheken">
      <binData>[base64-kodierter MAB-Datensatz weggelassen]</binData>
    </mdWrap>
  </dmdSec>
```

```
<amdSec>
  <techMD ID="METS42-techMD1" CREATED="2004-05-17T10:11:22">
    <mdWrap MDTYPE="NISOIMG" LABEL="">
      <xmlData>
        <mix:mix xmlns:mix="http://www.loc.gov/mix/"
          xsi:schemaLocation="http://www.loc.gov/mix/
            http://www.loc.gov/mix/mix.xsd">
          <mix:BasicImageParameters>
            <mix:Format>
              <mix:ByteOrder>little-endian</mix:ByteOrder>
              <mix:Compression>
                <mix:CompressionScheme>5</mix:CompressionScheme>
              </mix:Compression>
              <mix:PhotometricInterpretation>
                <mix:ColorSpace>2</mix:ColorSpace>
              </mix:PhotometricInterpretation>
              <mix:Segments>
                <mix:SegmentType>0</mix:SegmentType>
                <mix:StripOffsets>489428</mix:StripOffsets>
                <mix:StripOffsets>502650</mix:StripOffsets>
                  [zahlreiche Werte weggelassen]
                <mix:StripOffsets>196124968</mix:StripOffsets>
                <mix:RowsPerStrip>1</mix:RowsPerStrip>
                <mix:StripByteCounts>13222</mix:StripByteCounts>
                <mix:StripByteCounts>13123</mix:StripByteCounts>
                  [zahlreiche Werte weggelassen]
                <mix:StripByteCounts>19273</mix:StripByteCounts>
              </mix:Segments>
            </mix:Format>
          </mix:BasicImageParameters>
          <mix:ImageCreation>
              <mix:ImageProducer>Universitätsbibliothek Regensburg, Germany;
                Multimediazentrum</mix:ImageProducer>
              <mix:Host>
                <mix:HostComputer>PC with SCSI connection to scanner</mix:HostComputer>
                <mix:OperatingSystem>Windows</mix:OperatingSystem>
                <mix:OSVersion>2000</mix:OSVersion>
              </mix:Host>
              <mix:DeviceSource>reflection print scanner</mix:DeviceSource>
              <mix:ScanningSystemCapture>
                <mix:ScanningSystemHardware>
                  <mix:ScannerManufacturer>ProServ</mix:ScannerManufacturer>
                  <mix:ScannerModel>
                    <mix:ScannerModelName>TriAS A-0</mix:ScannerModelName>
                    <mix:ScannerModelNumber></mix:ScannerModelNumber>
                    <mix:ScannerModelSerialNo>052/2001</mix:ScannerModelSerialNo>
                  </mix:ScannerModel>
                </mix:ScanningSystemHardware>
                <mix:ScanningSystemSoftware>
                  <mix:ScanningSoftware>
                    Business Graphics Datentechnik ProView
                  </mix:ScanningSoftware>
                    <mix:ScanningSoftwareVersionNo>
                      3.84
                    </mix:ScanningSoftwareVersionNo>
                </mix:ScanningSystemSoftware>
              </mix:ScanningSystemCapture>
            </mix:ImageCreation>
```

```
          </mix:mix>
        </xmlData>
      </mdWrap>
    </techMD>
    <sourceMD ID="METS42-sourceMD1" CREATED="2004-05-17T10:14:03">
      <mdWrap MDTYPE="OTHER" OTHERMDTYPE="MAB2"
        LABEL="Datensatz im Maschinellen Austauschformat für Bibliotheken">
        <binData>[base64-kodierter MAB-Datensatz weggelassen]</binData>
      </mdWrap>
    </sourceMD>
  </amdSec>
  <fileSec>
    <fileGrp VERSDATE="2004-05-17T10:12:56" USE="master">
      <file ID="FILE126" MIMETYPE="image/tiff" SEQ="1" SIZE="196144284"
        CREATED="2004-05-17T10:11:22" CHECKSUM="49c00f62eddd389969a1eae49306325d"
        CHECKSUMTYPE="MD5" OWNERID="FILE126" GROUPID="GIDMETS42-1" USE="master">
        <FLocat LOCTYPE="URL" xlink:href="./master/12_Mapp_XI_57_du_002.tif"/>
      </file>
    </fileGrp>
  </fileSec>
  <structMap>
    <div>
      <div LABEL="Physische Struktur" TYPE="physical">
        <div ORDER="1" ORDERLABEL="1" LABEL="Seitenfolge" TYPE="page rank">
          <div ORDER="1" ORDERLABEL="1" LABEL="Seite 1" TYPE="page">
            <fptr FILEID="FILE126"/>
          </div>
        </div>
      </div>
      <div LABEL="Logische Struktur" TYPE="logical"/>
    </div>
  </structMap>
</mets>
```

4.2 Digitalisat eines ganzen Bandes mit Strukturinformationen

Am nächsten Beispiel, der Digitalisierung von Sturms „Deutschlands Flora in Abbildungen“ (Sturm 1797–1862), soll die Erschließung von Strukturmetadaten aus den Dateinamen gezeigt werden. Das ungebundene Werk handelt zahleiche Pflanzenarten auf jeweils ein bis drei Textseiten und einer farbigen Tafelseite ab. Die Einzelblätter wurden auf einem hochwertigen DIN-A3-Farbscanner gescannt. Da die Digitalisierung einer Einzelseite relativ aufwendig war, fiel die Kodierung von Strukturmetadaten in den Dateinamen kaum ins Gewicht. Beim Digitalisieren erfassten die Mitarbeiter in einem festen Muster die Seitenabfolge, die Pflanzenart sowie, ob es sich um eine Text- oder Tafelseite handelte. Diese Informationen konnten mit einem entsprechenden Parser für die Dateinamen automatisch extrahiert und in die METS-Datei aufgenommen werden. Angesichts der über 2700 Seiten des Werks bietet diese Erschließung eine große Erleichterung bei der Benutzung und die Art der automatischen Erschließung eine ungeheure Arbeitserleichterung.

Im folgenden sind repräsentative Ausschnitte aus der METS-Datei eines Bandes mit kurzen Erläuterungen dargestellt.

Die `<fileSec>` listet alle Dateien zusammen mit minimalen technischen Metadaten auf.

Die digitalen Master sind zu einer eigenen <fileGrp> zusammengefaßt:

```
<fileSec>
 <fileGrp VERSDATE="2004-05-17T11:56:27" USE="master">
  <file ID="FILE127" MIMETYPE="image/tiff" SEQ="1" SIZE="8701998"
    CREATED="2004-02-18T11:47:47"
    CHECKSUM="a771aef0dfd4db42195d4f2c56df9b33" CHECKSUMTYPE="MD5"
    OWNERID="FILE127" GROUPID="GIDMETS43-1" USE="master">
   <FLocat LOCTYPE="URL" xlink:href="master/Ap_Av_001_Apargia_alpina_Host_t1.tif"/>
  </file>
  <file ID="FILE128" MIMETYPE="image/tiff" SEQ="2" SIZE="8701792"
    CREATED="2004-02-18T11:47:51"
    CHECKSUM="64c70550ac73e1228eb9dbb193ee61e3" CHECKSUMTYPE="MD5"
    OWNERID="FILE128" GROUPID="GIDMETS43-2" USE="master">
   <FLocat LOCTYPE="URL" xlink:href="master/Ap_Av_002_Apargia_alpina_Host_t2.tif"/>
  </file>
  [restliche Dateien weggelassen]
 </fileGrp>
</fileSec>
```

Das oberste -Element in der <structMap> erhält an der Universitätsbibliothek Regensburg immer ein mit der physischen Struktur und ein mit der logischen Struktur des Dokuments. Die Erschließung der logischen Struktur kann beliebig tief sein. Hier wird nach den Pflanzenarten gegliedert, innerhalb derer jeweils die Text- und Tafelseiten aufgeführt und gekennzeichnet werden. Außerdem werden alle Abbildungen zusätzlich noch einmal in einem eigenen aufgeführt, da sie meist im Zentrum des Interesses an dem Werk stehen:

```
<structMap>
 <div>
  <div LABEL="Physische Struktur" TYPE="physical">
   <div ORDER="1" ORDERLABEL="1" LABEL="Seitenfolge" TYPE="page rank">
    <div ORDER="1" ORDERLABEL="1" LABEL="Seite 1" TYPE="page">
     <fptr FILEID="FILE127"/>
    </div>
    <div ORDER="2" ORDERLABEL="2" LABEL="Seite 2" TYPE="page">
     <fptr FILEID="FILE128"/>
    </div>
    <div ORDER="3" ORDERLABEL="3" LABEL="Seite 3" TYPE="page">
     <fptr FILEID="FILE129"/>
    </div>
    [weitere Dateien weggelassen]
   </div>
  </div>
  <div LABEL="Logische Struktur" TYPE="logical">
   <div ORDER="2" ORDERLABEL="2" LABEL="Pflanzenarten">
    <div LABEL="Apargia alpina Host" TYPE="page">
     <div ORDER="1" LABEL="Textseite 1" TYPE="page">
      <fptr FILEID="FILE127"/>
     </div>
     <div ORDER="2" LABEL="Textseite 2" TYPE="page">
      <fptr FILEID="FILE128"/>
     </div>
     <div ORDER="3" LABEL="Abbildungsseite 1" TYPE="page">
      <fptr FILEID="FILE129"/>
     </div>
    </div>
```

```
    <div LABEL="Apargia crocea Willd" TYPE="page">
     <div ORDER="1" LABEL="Textseite 1" TYPE="page">
      <fptr FILEID="FILE130"/>
     </div>
     <div ORDER="2" LABEL="Textseite 2" TYPE="page">
      <fptr FILEID="FILE131"/>
     </div>
     <div ORDER="3" LABEL="Abbildungsseite 1" TYPE="page">
      <fptr FILEID="FILE132"/>
     </div>
    </div>
    [weitere Arten weggelassen]
   </div>
   <div ORDER="3" ORDERLABEL="3" LABEL="Abbildungen">
    <div LABEL="Apargia alpina Host" TYPE="page">
     <fptr FILEID="FILE129"/>
    </div>
    <div LABEL="Apargia crocea Willd" TYPE="page">
     <fptr FILEID="FILE132"/>
    </div>
    <div LABEL="Apargia dubia Hoppii" TYPE="page">
     <fptr FILEID="FILE135"/>
    </div>
    [weitere Arten weggelassen]
   </div>
  </div>
 </div>
</structMap>
```

4.3 Erzeugung eines Inhaltverzeichnisses

Die METS-Datei des letzten Abschnitts erweiterte ich die um die Metadaten für ein Derivat, nämlich jpg-Dateien für die Präsentation. Aus diesem METS-Dokuments erstellte ich dann durch XSL Transformation eine HTML-Seite, die als Inhaltsverzeichnis und Einstieg für eine CD-ROM dienen kann. Mit dem XSL-Stylesheet werden alle Versionen einer Datei des digitalen Dokuments, d. h. Master und Derivate, aufgelistet. Die logische und die physische Struktur des Dokuments werden übersichtlich in ihrer Hierarchie dargestellt. Es werden außerdem alle Dateien aller Versionen des Dokuments zusammen mit wichtigen technischen Angaben aufgelistet. Dabei werden auch Berechnungen vorgenommen, wie die Angabe der Anzahl der Dateien. Die Bilddateien sind über Hyperlinks direkt anzuspringen.

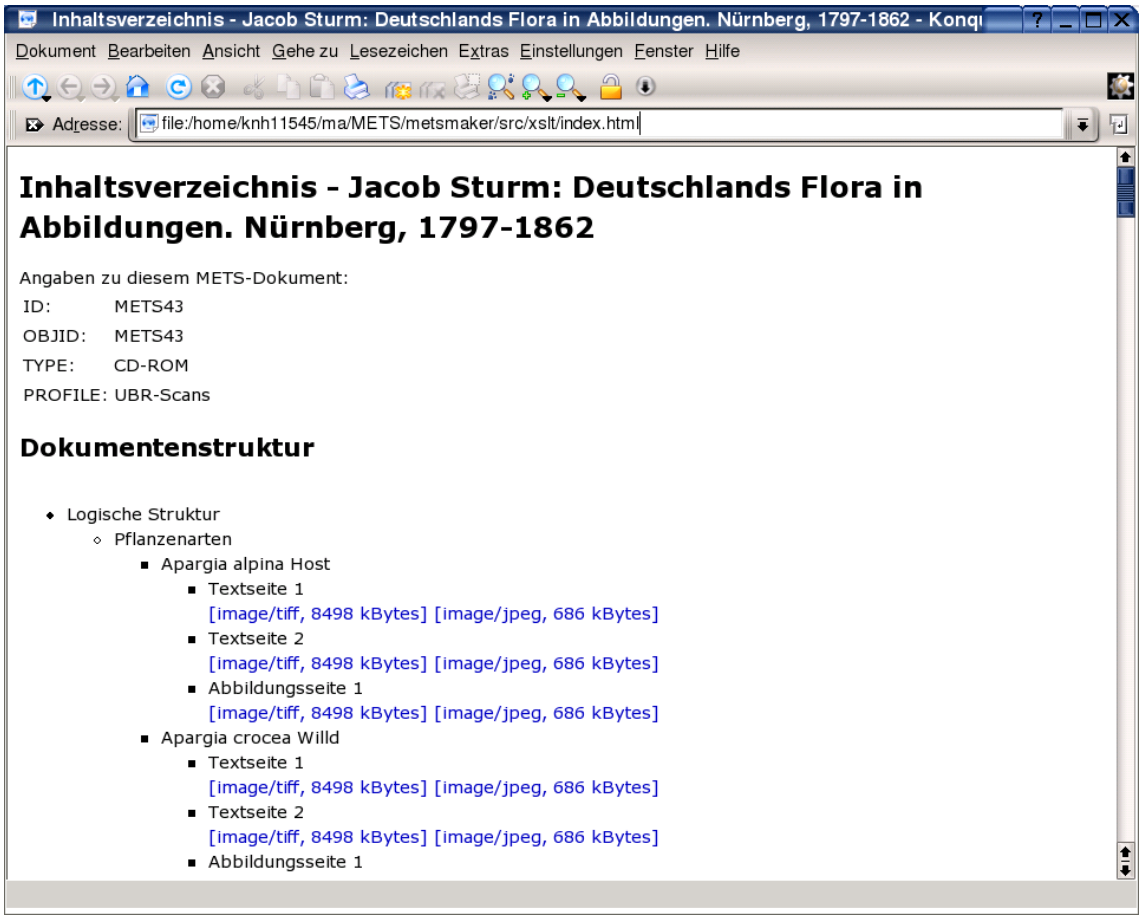

Abb. 4.1: Ansicht einer aus einem METS-Dokumente generierten HTML-Datei im Webbrowser: Seitenkopf mit Angaben zum METS-Dokument (Meta-Metadaten) und erster Teil der Dokumentenstruktur.

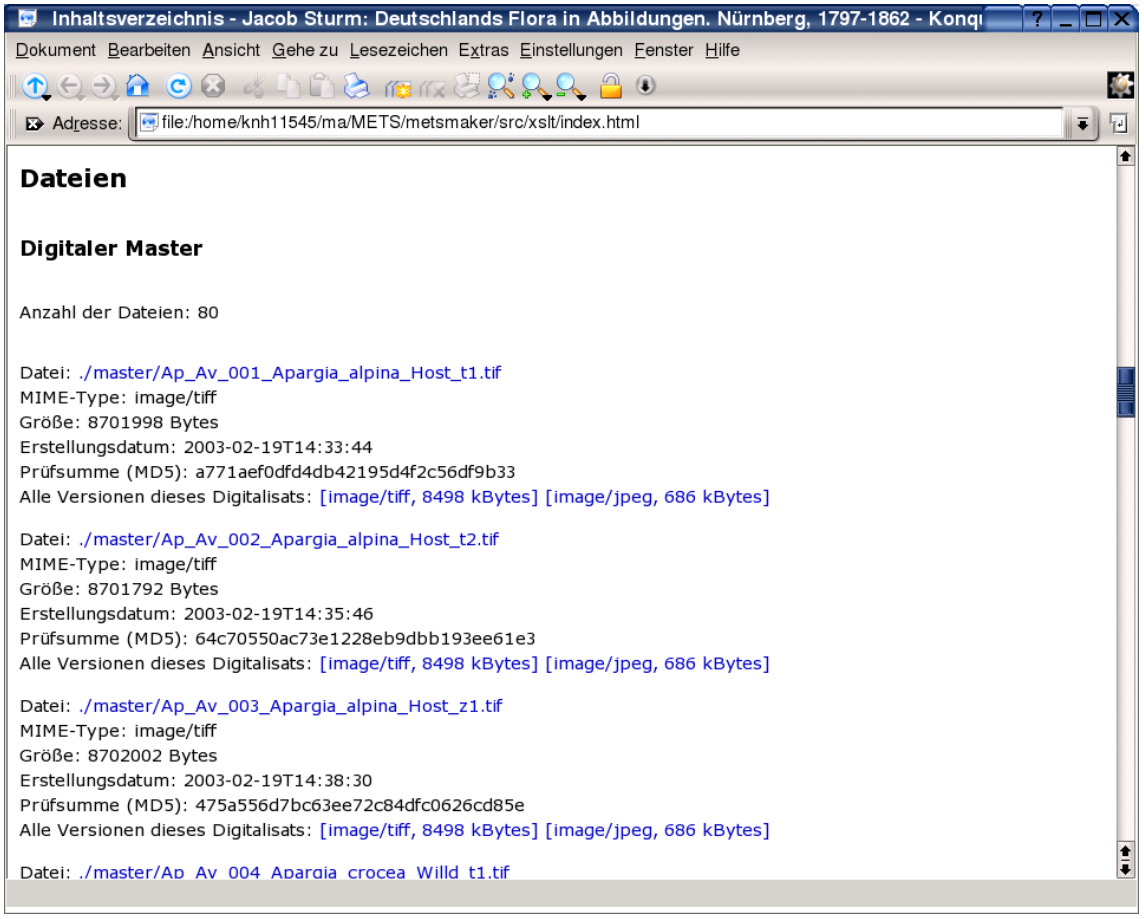

Abb. 4.2: Ansicht einer aus einem METS-Dokumente generierten HTML-Datei im Webbrowser: Erster Teil der Dateiliste.

5

Diskussion und Bewertung

Zum Abschluß dieser Arbeit möchte ich noch kurz diskutieren und bewerten, wie sich die erarbeiteten Konzepte bewährten und wie sich die Implementierung gestaltete.

In dieser Arbeit sollte ein Metadaten-Rahmenwerk für die Digitalisierung an der Universitätsbibliothek Regensburg konzipiert und prototypisch implementiert werden. Es sollten deskriptive, administrative und strukturelle Metadaten erfaßt werden können, um eine geeignete Basis für die digitale Langzeitarchivierung und für die Erstellung von Präsentationsformen zu schaffen. Das Metadatenformat mußte sich außerdem am Referenzmodell des Open Archival Information Systems (OAIS 2002) orientieren.

Die Suche nach einem geeigneten, standardisierten Metadatenformat für Digitalisate ergab eine erstaunliche Konvergenz der Entwicklungen von bibliothekarischer Seite hin zu zwei Entwicklungen, dem Ansatz von OCLC/RLG (PMFWG 2002) und zum Metadata Encoding and Transmission Standard METS (METS 2003). Da von diesen bisher nur für den METS auch eine technische Spezifikation vorlag und er damit sofort einsatzfähig war, fiel die anschließende Wahl des METS leicht.

Die Entscheidung für den METS bereute ich bisher nicht. Die Beschäftigung mit dem Standard erwies sich vor allem anfangs als recht anspruchsvoll, obwohl ich bisher nur einen Teil seiner Möglichkeiten nutzte. Allerdings blieb bisher auch kein Wunsch offen und der Standard erwies sich mehr als einmal als wohldurchdacht. METS-Dokumente können als universelles Austausch- und Archivierungsformat dienen, wobei die Inhalte der digitalen Dokumente zusammen mit den Metadaten in einer Datei vereinigt werden können.

Bei der Planung des Metadaten-Rahmenwerks erwies sich ein System zur Vergabe von Persistent Identifiern (PIDs) als notwendig. PIDs werden für CD-ROMs, METS-Dokumente und Dateien benötigt. Beim Übergang von der Speicherung der digitalen Dokumente auf CD-ROMs zu einem Online-System können die CD-ROM-PIDs entfallen. Das implementierte System zur Vergabe von PIDs kann jederzeit leicht zu einem URN-System mit Resolver weiterentwickelt werden. Neben der PID-Vergabe dient es auch der Authentizitätsprüfung für die digitalen Dokumente.

Die neuartige Erschließung ließ sich problemlos in den Geschäftsgang integrieren. Der Geschäftsgang konnte in diesem Zusammenhang sogar noch etwas verbessert werden, indem die Geschäftsprozesse nach der Digitalisierung eindeutiger geregelt wurden und ein Dateiserver dafür mit einbezogen wurde. Wie sich das neue Verfahren letztlich im täglichen Routinebetrieb bewähren wird, bleibt noch abzuwarten.

Ich programmierte prototypisch ein Java-basiertes Erschließungswerkzeug, das komplett modular gestaltet ist, um größtmögliche Flexibilität für die Verarbeitung unterschiedlicher Dokumente und Dateiformate sowie für zukünftige Erweiterungen zu erreichen. Das

Programm ist plattformunabhängig und als Client-Applikation konzipiert. Es soll die Erschließung der Digitalisate soweit wie möglich automatisieren. Im einfachsten Fall genügt die manuelle Zuordnung gerätespezifischer und bibliographischer Metadaten zum Digitalisat, um die Erschließung dann automatisch ablaufen zu lassen. Die Arbeiten gediehen soweit, daß damit erste Dokumente erschlossen werden konnten.

Daneben wird die Schaffung einer graphischen Benutzeroberfläche die Bedienung noch deutlich erleichtern sowie zusätzliche Möglichkeiten der Erschließung schaffen.

Die Kodierung von Strukturinformationen in den Dateinamen hat sich als sehr effiziente und wirkungsvolle Methode der Metadatenerfassung herausgestellt. Die Übernahme der Informationen mit geeigneten Parsern für die Dateinamen ist deshalb eine der Kernfunktionen des von mir programmierten Erschließungswerkzeugs. Das DFG-Papier „Die Erschließung und Bereitstellung digitalisierter Drucke“ (DFG 2002) führt in Anlage I, Punkt 2 einen Vorschlag für die Namenskonvention für digitale Medieneinheiten auf. Dieser Vorschlag erwies sich unter den Rahmenbedingungen an der Universitätsbibliothek Regensburg als nicht anwendbar. Unabhängig davon ist eine persistente Adressierung, wie in dieser Arbeit vorgeschlagen, jederzeit möglich.

Die in METS-Dokumenten kodierten Metadaten können der flexiblen Erzeugung von Präsentationsformen der digitalen Dokumente, etwa mittels XSLT, dienen. Dadurch ließ sich die Benutzbarkeit der Digitalisate deutlich verbessern.

Die Ergebnisse dieser Arbeit lassen sich sicherlich auch von anderen Institutionen nutzen, die in ähnlicher Weise digitalisieren. Es ist geplant, die Programme alleine schon für den Einsatz im eigenen Hause zu erweitern und zu verbessern. Der Quelltext der Programme soll unter einer Open-Source-Lizenz freigegeben werden.

Literaturverzeichnis

Abele 2004 ABELE, Heinrich: Der Tübinger Internet Multimedia Server "timms„ – das multimediale Forschung-und-Lehre-Server-System an der Universität Tübingen. In: *Praxis der Informationsverarbeitung und Kommunikation (PIK)* 27 (2004), Nr. 1, S. 3–9. – ISSN 0930-5157

Aschoff u. a. 2004 ASCHOFF, Christian ; HÄNISCH, Till ; GROSSMANN, Hans P.: Das DLmeta-Modell als Grundlage für verteilte Multi-Media-Archive. In: *Praxis der Informationsverarbeitung und Kommunikation (PIK)* 27 (2004), Nr. 1, S. 27. – ISSN 0930-5157

AUDIOMD 2002 : *Audio Technical Metadata Extension Schema*. August 22 2002. – URL `http://lcweb2.loc.gov/mets/Schemas/AMD.xsd`. – Zugriffsdatum: 05.05.2004

Berners-Lee u. a. 1998 BERNERS-LEE, T. ; FIELDING, R. ; MASINTER, L.: *Uniform Resource Identifiers (URI): Generic Syntax*. Request For Comments (RFC) 2396. August 1998. – URL `http://www.ietf.org/rfc/rfc2396.txt`. – Zugriffsdatum: 06.05.2004

Berners-Lee u. a. 1994 BERNERS-LEE, T. ; MASINTER, L. ; MCCAHILL, M.: *Uniform Resource Locators (URL)*. Request For Comments (RFC) 1738. Dezember 1994. – URL `http://www.ietf.org/rfc/rfc1738.txt`. – Zugriffsdatum: 06.05.2004

Cedars 2002 : *Cedars: CURL Exemplars in Digital Archives*. Website. 2002. – URL `http://www.leeds.ac.uk/cedars/`. – Zugriffsdatum: 30.04.2004

Cedars-Metadaten 2002 *Metadata for Digital Preservation: the Cedars Project Outline Specification*, Januar 22 2002. – URL `http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html`. – Zugriffsdatum: 30.04.2004

Cooper und Garcia-Molina 2002 COOPER, Brian F. ; GARCIA-MOLINA, Hector: Peer-to-peer data trading to preserve information. In: *ACM Transactions on Information Systems* 20 (2002), Nr. 2, S. 133–170

Cullen u. a. 2000 CULLEN, Charles T. ; HIRTLE, Peter B. ; LEVY, David ; LYNCH, Clifford A. ; ROTHENBERG, Jeff: *Authenticity in a Digital Environment*. Washington, DC : Council on Library and Information Resources (CLIR), April 2000 (CLIR Reports). – 1–84 S. – URL `http://www.clir.org/pubs/reports/pub92/pub92.pdf`. – Zugriffsdatum: 20.04.2004. – ISBN 1-887334-77-7

Cundiff 2003 CUNDIFF, Morgan: *NISO Metadata for Images in XML (NISO MIX). Draft Version 0.2, April 11, 2003*. April 2003. – URL `http://www.loc.gov/standards/mix/mix.xsd`. – Zugriffsdatum: 08.03.2004

Day 2001 DAY, Michael: Metadata for Digital Preservation: A Review of Recent Developments. In: *Lecture Notes in Computer Science* 2163 (2001), S. 161–172. – URL http://www.springerlink.com/openurl.asp?genre=article&issn=0302-9743&volume=2163&spage=161. – Zugriffsdatum: 20.04.2004. – ISSN 0302-9743

Devlin u. a. 2004 DEVLIN, Kate ; CHALMERS, Alan ; REINHARD, Erik: Displaying digitally archived images. In: *Proceedings of the IS&T Archiving Conference*. San Antonio, Texas : The Society for Imaging Science and Technology, April 20–23, 2004

DFG 2002 *Die Erschließung und Bereitstellung digitalisierter Drucke. Vorschläge des Unterausschusses für Kulturelle Überlieferung*, Oktober 2002. – 1–43 S. – URL http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/download/konzept_digitale_drucke.pdf. – Zugriffsdatum: 19.04.2004

DLF & CLIR 2000 *Guides to Quality in Visual Resource Imaging*, Juli 2000. – URL http://www.rlg.org/visguides/. – Zugriffsdatum: 27.04.2004

DLmeta 2000 : *DLmeta Initiative*. Website. 2000. – URL http://www.dlmeta.de/. – Zugriffsdatum: 28.04.2004

DLmeta-DTD 2000 : *Document Type Definition (DTD) des DLmeta-Datenmodells*. 2000. – URL http://www.dlmeta.de/jdlmeta/dtd/index.jsp. – Zugriffsdatum: 28.04.2004

Dobratz u. a. 2001 DOBRATZ, Susanne ; LIEGMANN, Hans ; TAPPENBECK, Inka: Langzeitarchivierung digitaler Dokumente. In: *ZfBB* 48 (2001), Nr. 6, S. 327–332

Dublin Core 2003 : *Dublin Core Metadata Element Set, Version 1.1: Reference Description*. Juni 2 2003. – URL http://www.dublincore.org/documents/dces/. – Zugriffsdatum: 30.04.2004

Ebind 2002 POLLOCK, Alvin: *Electronic Binding DTD (Ebind)*. Website. April 4, 2002. – URL http://sunsite.berkeley.edu/Ebind/. – Zugriffsdatum: 27.04.2004

Ebind-DTD 1996 POLLOCK, Alvin: *Electronic Binding DTD (Ebind). Version 0.1*. Juli 25, 1996. – URL http://sunsite.berkeley.edu/Ebind/ebind.dtd. – Zugriffsdatum: 27.04.2004

Fielding u. a. 1999 FIELDING, R. ; GETTYS, J. ; MOGUL, J. ; FRYSTYK, H. ; MASINTER, L. ; LEACH, P. ; BERNERS-LEE, T.: *Hypertext Transfer Protocol – HTTP/1.1*. Request For Comments (RFC) 2616. Juni 1999. – URL http://www.ietf.org/rfc/rfc2616.txt. – Zugriffsdatum: 06.05.2004

Freed und Borenstein 1996 FREED, N. ; BORENSTEIN, N.: *Multipurpose Internet Mail Extensions (MIME) Part One: Media Types*. Request For Comments (RFC) 2046. November 1996. – URL http://www.ietf.org/rfc/rfc2046.txt. – Zugriffsdatum: 12.05.2004

Frey 2000 FREY, Franziska: File Formats for Digital Masters / Digital Library Federation and Council on Library and Information Resources. URL http://www.rlg.org/visguides/visguide5.html. – Zugriffsdatum: 27.04.2004, Juli 2000. – Richtlinie

Gieselmann 2004 GIESELMANN, Hartmut: Blaues Gedächtnis. Professionelle Datensicherung der nächsten Generation. In: *c't – Zeitschrift für Computertechnik* 2004 (2004), März, Nr. 6, S. 196–200

Hakala 2001 HAKALA, J.: *Using National Bibliography Numbers as Uniform Resource Names*. Request For Comments (RFC) 3188. Oktober 2001. – URL http://www.ietf.org/rfc/rfc3188.txt. – Zugriffsdatum: 06.05.2004

Hammen und Slotta 2004a HAMMEN, Ralf ; SLOTTA, Alexander: Das Uniform-Resource-Name (URN) Management Der Deutschen Bibliothek für Online-Hochschulschriften / Die Deutsche Bibliothek. URL http://www.persistent-identifier.de/?link=311. – Zugriffsdatum: 05.04.2004, Februar 2004. – Forschungsbericht

Hammen und Slotta 2004b HAMMEN, Ralf ; SLOTTA, Alexander: EPICUR: Uniform Resource Name (URN) - Strategie Der Deutschen Bibliothek / Die Deutsche Bibliothek. URL http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:1111-2003121811. – Zugriffsdatum: 05.04.2004, März 2004. – Forschungsbericht. URN: urn:nbn:de:1111-2003121811

Hammen und Slotta 2004c HAMMEN, Ralf ; SLOTTA, Alexander: Informationen für die Uniform Resource Name (URN)-Verwaltung und -Geschäftsverteilung für Universitätsbibliotheken / Die Deutsche Bibliothek. URL http://www.persistent-identifier.de/?link=312. – Zugriffsdatum: 05.04.2004, Februar 2004. – Forschungsbericht

JCDL'02 2002 CORFIELD, Anthony ; DOVEY, Matthew ; MAWBY, Richard ; TATHAM, Colin: JAFER ToolKit Project – Interfacing Z39.50 and XML. In: MARCHIONINI, Gary (Hrsg.) ; HERSH, William (Hrsg.): *JCDL 2002: Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*. Portland, Oregon : Association for Computing Machinery, Juli 14–18 2002, S. 289–290. – ISBN 1-58113-513-0

Kossel 2003 KOSSEL, Axel: Gute Karten für alle. In: *c't – Zeitschrift für Computertechnik* 2003 (2003), April, Nr. 8, S. 3

LoC 2004 : *Extension Schemas for the Metadata Encoding and Transmission Standard*. Website. 2004. – URL http://lcweb.loc.gov/rr/mopic/avprot/metsmenu2.html. – Zugriffsdatum: 05.05.2005

MAB2 1995 *MAB2. Maschinelles Austauschformat für Bibliotheken*. Frankfurt am Main, 1995

MABxml Website 2004 : *MABxml*. Website. 2004. – URL http://www.ddb.de/professionell/mabxml.htm. – Zugriffsdatum: 19.03.2004

McDonough 2003 MCDONOUGH, Jerome: METS Profile Documentation. Version 1.0 / Library of Congress. Washington, D. C., 2003. – Datei im Rich-Text-Format. – URL http://www.loc.gov/standards/mets/profile_docs/METS.profile.requirements.rtf. – Zugriffsdatum: 02.03.2004

METS 2003 : *Metadata Transmission and Encoding Standard (METS)*. Website. Juni 2003. – URL http://www.loc.gov/standards/mets/

METS-Java-Toolkit 2004 ABRAMS, Stephen: *METS Java Toolkit, Version 1.3*. Website des Software-Projekts. 2004. – URL http://hul.harvard.edu/mets/. – Zugriffsdatum: 11.03.2004

METS Schema 1.3 2003 : *METS Schema 1.3*. Mai 8, 2003. – URL http://www.loc.gov/standards/mets/mets.xsd. – Zugriffsdatum: 02.03.2004

MIX-Website 2003 : *NISO Technical Metadata for Digital Still Images in XML*. Website. Juni 2003. – URL http://www.loc.gov/standards/mix/. – Zugriffsdatum: 14.04.2004

MOA2 2001 : *The Making of America II*. Website. Oktober 10, 2001. – URL http://sunsite.berkeley.edu/MOA2/. – Zugriffsdatum: 27.04.2004

MOA2-DTD 2000 MCDONOUGH, Jerome P.: *The Making of America II Document Type Definition. Version 2.0 (BETA 1.3)*. März 13, 2000. – URL http://sunsite.berkeley.edu/moa2/papers/moa2dtd2.htm. – Zugriffsdatum: 27.04.2004. – Die DTD ist eingebettet in ein HTML-Dokument

Moats 1997 MOATS, R.: *URN Syntax*. Request For Comments (RFC) 2141. Mai 1997. – URL http://www.ietf.org/rfc/rfc2141.txt. – Zugriffsdatum: 06.05.2004

NEDLIB 2000 : *Project NEDLIB – Networked European Deposit Library*. Website. 2000. – URL http://www.kb.nl/coop/nedlib/. – Zugriffsdatum: 30.04.2004

NEDLIB-Metadaten 2000 LUPOVICI, Catherine ; MASANÈS, Julien: *Metadata for long term-preservation*. Den Haag: , Juli 2000. – URL http://www.kb.nl/coop/nedlib/results/NEDLIBmetadata.pdf. – Zugriffsdatum: 30.04.2004

NLA 1999 National Library of Australia (Veranst.): *Preservation Metadata for Digital Collections*. Oktober 15 1999. – URL http://www.nla.gov.au/preserve/pmeta.html. – Zugriffsdatum: 30.04.2004

NORM ISO 12234-2:2001 2001 *ISO 12234-2:2001. Electronic still-picture imaging – Removable memory – Part 2: TIFF/EP image data format*, 2001. – 1–60 S

NORM ISO 14721:2003 2003 *ISO 14721:2003. Space data and information transfer systems – Open archival information system – Reference model*, 2003. – 1–156 S. – Auch verfügbar als OAIS (2002)

NORM Z39.87-2002 2002 National Information Standards Organization and AIIM International (Veranst.): *Data Dictionary – Technical Metadata for Digital Still Images. Draft Standard for Trial Use*. 2002. – URL http://www.niso.org/standards/resources/Z39_87_trial_use.pdf. – Zugriffsdatum: 08.03.2004

OAIS 2002 *Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1. Blue Book. Issue 1.*, Januar 2002. – URL http://www.ccsds.org/documents/650x0b1.pdf. – Zugriffsdatum: 10.12.2003. – Wurde als NORM ISO 14721:2003 zum Standard erhoben

PMFWG 2001 OCLC/RLG Preservation Metadata Framework Working Group (Veranst.): *Preservation Metadata for Digital Objects: A Review of the State of the Art*. Januar 31 2001. – URL http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf. – Zugriffsdatum: 30.04.2004

PMFWG 2002 OCLC/RLG Preservation Metadata Framework Working Group (Veranst.): *Preservation Metadata and the OAIS Information Model. A Metadata Framework to Support the Preservation of Digital Objects*. Juni 2002. – URL http://www.oclc.org/research/projects/pmwg/pm_framework.pdf. – Zugriffsdatum: 30.04.2004

Reich 2002 REICH, Victoria A.: Diffused Knowledge Immortalizes Itself. The LOCKSS Program. In: *High Energy Physics Libraries Webzine* 7 (2002), Januar 01. – URL http://library.cern.ch/HEPLW/7/papers/1/;http://eprints.rclis.org/archive/00000447/02/diffused_knowledge.pdf. – Zugriffsdatum: 20.04.2004

Schroeder 2003 SCHROEDER, Kathrin: Persistent Identifiers im Kontext von Metadaten. In: *ZfBB* 50 (2003), Nr. 4, S. 205–209

Schulzki-Haddouti 2003a SCHULZKI-HADDOUTI, Christiane: Bedingt gerichtsfest. Studie testet Beweiskraft digitaler Signaturen. In: *c't – Zeitschrift für Computertechnik* 2003 (2003), Dezember, Nr. 23, S. 40. – URL http://www.heise.de/ct/03/23/040/default.shtml

Schulzki-Haddouti 2003b SCHULZKI-HADDOUTI, Christiane: IT-Sicherheit für den Mittelstand. Das BSI forciert Beratung in Sicherheitsfragen. In: *c't – Zeitschrift für Computertechnik* 2003 (2003), Juni, Nr. 12, S. 42

Spanier 2002 SPANIER, Kurt: DLmeta2002 – Die Evolution eines Datenmodells. In: *Benutzerinformationen des ZDV* 2002 (2002), Nr. 7–12, S. 8–13. – URL http://www.uni-tuebingen.de/zdv/bi/bi02/bi029e2-dlmeta.html. – Zugriffsdatum: 28.04.2004. – ISSN 1430-5577

Spanier 2004 SPANIER, Kurt: DLmeta2002 – Evolution und Einsatz eines Meta-Datenmodells. In: *Praxis der Informationsverarbeitung und Kommunikation (PIK)* 27 (2004), Nr. 1, S. 32–36. – ISSN 0930-5157

Sturm 1797–1862 STURM, Jacob: *Deutschlands Flora in Abbildungen*. Nürnberg, 1797–1862

TIFF 6.0 1992 *TIFF, Revision 6.0*. Final – June 3, 1992, 1992. – URL http://partners.adobe.com/asn/developer/pdfs/tn/TIFF6.pdf. – Zugriffsdatum: 08.03.2004

Werkmeister und Bognar 2004 WERKMEISTER, Walter ; BOGNAR, Istvan: Die Lokale Medienbibliothek (LMB) der Universitätsbibliothek Tübingen. In: *Praxis der Informationsverarbeitung und Kommunikation (PIK)* 27 (2004), Nr. 1, S. 21–26. – ISSN 0930-5157

XML 2004 YERGEAU, François (Hrsg.) ; BRAY, Tim (Hrsg.) ; PAOLI, Jean (Hrsg.) ; SPERBERG-MCQUEEN, C. M. (Hrsg.) ; MALER, Eve (Hrsg.): *Extensible Markup Language (XML) 1.0. W3C Recommandation*. Website. Februar 4 2004. – URL http://www.w3.org/TR/REC-xml

XML-Schema 2001 : *XML-Schema 1.0. W3C Recommendation*. Website. Mai 2, 2001. – URL http://www.w3.org/XML/Schema

XPath 1999 CLARK, James ; DEROSE, Steve: *XML Path Language (XPath). Version 1.0*. Website. November 16, 1999. – URL http://www.w3.org/TR/xpath. – Zugriffsdatum: 16.05.2004

XSLT 1999 : *XSL Transformations (XSLT). Version 1.0. W3C Recommendation*. Website. November 16, 1999. – URL `http://www.w3.org/TR/xslt`

Danksagung

Ich möchte mich vor allem bei Dr. Albert Schröder herzlich bedanken, der mir die Bearbeitung des Themas als Teil meiner Tätigkeit für das Projekt Bayerische Landesbibliothek Bayern ermöglichte. Er stand mir immer als kundiger Diskussionspartner zur Verfügung und setzte sich bereitwillig und gründlich mit dem Manuskript auseinander.

Die Mitarbeiterinnen des Multimediazentrums der Universitätsbibliothek Regensburg, allen voran Gabriele Gerber, M. A., danke ich für ihre immer freundliche Zusammenarbeit und geduldige Erläuterung des Geschäftsgangs bei der Digitalisierung.

Stephen Abrams von der Harvard University Library stellte mir auf Nachfrage sofort sein Java METS Toolkit unter der GNU General Public License zur Verfügung.

Dr. Kurt Spanier vom Zentrum für Datenverarbeitung der Universität Tübingen gab mir bereitwillig Auskunft zum Stand der Arbeiten am DLmeta-Standard.

Christof Mainberger vom Bibliotheksservice-Zentrum Baden-Württemberg sandte mir freundlicherweise Teile eines DFG-Antrags zum Aufbau eines Metadaten-Repository im Rahmen des Verteilten Dokumentenservers zu.

Bei Marc Reymann, M. A. bedanke ich mich für Informationen zur URN-Vergabe und Metadatenstruktur beim Opus-System in Bayern[1].

Zahlreiche weitere Personen, vor allem aus dem Kollegenkreis in Regensburg, standen mir mit Ratschlägen und Hilfe bei der Erstellung dieser Arbeit zur Seite. Ich möchte mit hiermit bei ihnen allen herzlich bedanken.

Meine Frau und meine Kinder hatten während der Bearbeitungszeit dieser Masterarbeit nicht viel von mir. Vielen Dank für Euer Verständnis, jetzt wird es besser!

[1] Siehe `http://www.opus-bayern.de/`